Reciprocity, Institutions, and Cooperation

Dissertation
for the Faculty of Economics, Business Administration and Information Technology of the University of Zurich

to achieve the title of
Doctor of Philosophy
in Economics

presented by

Silvia Christine Verena Grätz
from Germany

approved in July 2012 at the request of

Prof. Dr. Armin Schmutzler
Prof. Dr. Michelle Sovinsky

The Faculty of Economics, Business Administration and Information Technology of the University of Zurich hereby authorizes the printing of this Doctoral Thesis, without thereby giving any opinion on the views contained therein.

Zurich, July 18, 2012

Chairman of the Doctoral Committee: Prof. Dr. Dieter Pfaff

Contents

Acknowledgement

I am very grateful to several people without whose support, contributions, and guidance this thesis would not have been completed.

I am indebted to my advisor Armin Schmutzler for his professional advice and guidance of my research and Ph.D studies, and for his belief in my research ideas. I also wish to thank Michelle Sovinsky for giving me many inspiring and insightful comments to improve my research and for sparking my interest in empirical research. Moreover, I would like to express my gratitude to Nick Netzer for his tremendous encouragement, guidance, and motivation throughout my struggle with Chapter 1.

I especially thank my co-author Donja Darai, whose positive attitude and enthusiasm I always enjoyed and which made working with her not only fruitful, but also lots of fun.

I am thankful to some former and present colleagues in Zurich: Simon Alder, Dennis Gärtner, Daniel Halbheer, Andreas Hefti, Stefan Jönsson, Arnd Heinrich Klein, Igor Letina, Konrad Mierendorf, Frédéric Schneider, and Christoph Winter, for many stimulating discussions, for the motivating atmosphere at our department, and for all the fun we had during the last four years. They all made my doctoral studies a great experience.

Last but not least, I wish to thank my family, my parents Christa and Klemens Grätz for their care and invaluable support during my life, as well as my friends for allowing me to always count on them. Foremost, I thank Kristoph Steikert for his love, patience and faith in me, and for his full support in every possible way I can think of.

Zurich, 2012 Silvia Grätz

Introduction

Traditional economic theory assumes that agents have selfish preferences about material payoffs. They do not care whether their actions affect others, and act as maximizers of their own material well-being. This assumption leads to strong theoretical predictions about economic decision making. However, as evidence from experimental and empirical research in psychology and economics shows, observed behavior is in conflict with theoretically predicted behavior, suggesting that individuals are guided by more than just selfish preferences. For their well-being intentions behind their (and possibly others') actions as well as the consequences for the outcomes of others may also matter. Hence, people have concerns for others and exhibit social preferences.
The present work is a collection of three essays providing a theoretical analysis and empirical evidence of the implications of interdependent social preferences for game-theoretic situations. The first essay focuses on the theory of implementation and studies implementation problems when agents are inclined to respond to the behavior of others in a reciprocal way. The second essay explores the formation of cooperative behavior in an empirical analysis of a variant of the prisoner's dilemma game and highlights the relevance and informational value of communication. The last essay continuous the empirical investigation of cooperative behavior and emphasizes the potential of people's physical attractiveness to elicit cooperation.

Chapter one studies the problem of a planner who faces a number of agents and wishes to implement a socially optimal outcome. The implementation problem arises due to the asymmetric information between the planner and the agents. While the agents collectively know their true preferences over the set of outcomes, the planner does not. Instead, the planner relies on the agents' individual reports in form of messages, that contain information about their preferences. However, an agent's message can be a strategic lie in order to influence the outcome to her advantage. An illustrative example of such a scenario is an auction in which agents may strategically select their bids, balancing prices and winning probabilities. The design or

rules of the institution, the mechanism, through which the agents interact therefore has a major impact on the agents' strategic behavior and on the outcomes of the interaction. The purpose of an auction, for instance, is to make sure that the winner is the one who values the auctioned object the most. Accordingly, a mechanism should induce the agents to reveal their true preferences over outcomes.
The classic approach to implementation assumes that agents are motivated solely by the pursuit of self-interest (selfish preferences). Contrary to most of the existing literature on implementation, the present essay assumes that agents exhibit intention-based reciprocity preferences. They are willing to sacrifice their own material well-being in order to either punish behavior by others that they perceive as unkind, or to reward behavior by others that they perceive as kind. The equilibrium concept is a Nash fairness equilibrium and a notion of Nash fairness implementation is introduced, which captures reciprocity motives by the agents.
The first part of the analysis focuses on the canonical mechanism, that has been used for Nash implementation. The results show that the implementation problem might not be solvable with the canonical mechanism when agents have reciprocity preferences. A decisive role plays the agent's individual willingness to trade-off material payoffs and reciprocal kindness. In three abstract settings, in which the agents' preferences are structured differently, the potential but also the limits of the canonical mechanism are shown. In the first two settings the canonical mechanism either has no equilibrium, or many unwanted equilibria if the agents' concern for reciprocity is large enough (exceeding a certain threshold). In the third setting, however, it can be shown that the canonical mechanism yields the intended outcome, irrespective of the agents' willingness to trade-off material payoffs and reciprocal kindness, in which case the social choice function admits a "psychologically robust" implementation.
In the second part of the analysis the psychological robustness is addressed explicitly and analyzed for general mechanisms. The main result establishes that every Nash equilibrium of any mechanism, in which a single agent cannot affect the equilibrium outcome, is also a psychologically robust equilibrium. The problem is, however, that a psychologically robust equilibrium, which has the desired outcome in one state, continuous to be an equilibrium for all other states. This, in turn, implies that the equilibrium outcome need not be unique. Hence, as soon as psychological robustness is obtained, there are many psychologically robust equilibria which may not have the desired outcome.

Chapters two and three are motivated by the observation that people cooperate in the absence of strategic or material incentives to do so. The investigation of cooperative behavior constitutes a major topic in economics, and the well-known prisoner's dilemma game has become the classic economic example to demonstrate non-cooperative behavior. Two agents face a dilemma in which, independent of the other's action, each agent is better off by defection than by cooperation. However,

the outcome obtained when both defect is worse for each agent than the outcome obtained if both had cooperated. Traditional economic theory predicts no cooperation, though it leads to a solution in which both agents end up worse off than if both act contrary to self-interest.

In chapter two, data from the British television show "Golden Balls" is used to analyze people's behavior in a prisoner's dilemma, in which defection is a weakly dominant strategy. The game show provides an environment with extraordinarily high stakes, face-to-face communication between players and pre-play interactions. In the pre-play, contestants accumulate stakes and select their partner for the prisoner's dilemma, in which they play to eventually share a jackpot. This jackpot, on average, amounts to £13 000, and ranges from £3 to £100 150. Material incentives are therefore high-powered and might be expected to offset any other concerns like fairness or reciprocity.
The main results are a unilateral cooperation rate of 54% and a mutual cooperation rate of 33%. Communication, both verbal and non-verbal, and stake size have a major impact on cooperative behavior. In particular, promises and handshakes are used to initiate cooperation from the opponent. But, while a mere promise in combination with a handshake increases a contestant's likelihood to cooperate, a handshake alone increases the likelihood to defect. Although one might expect that handshakes serve as a positive commitment device, contestants lie when shaking hands, and use them as an instrument to manipulate the opponent's attitude towards cooperation. Also, a negative correlation between stake size as well as expected stake size and cooperation is observed.
Further, there is a strong link between contestants' behavior in the pre-play and the behavior in the prisoner's dilemma. Contestants who, from a purely material perspective, should have been voted off the game in the pre-play, show a higher propensity to cooperate. Here, cooperation can be interpreted as a result of reciprocal behavior: contestants who owe their survival of the pre-play to their final opponent, although their survival involves no material gain, are willing to cooperate in order to reward the kind behavior by the opponent. The data also offers the opportunity to test for consistent behavior of contestants in the sense that they act in accord with a pre-announcement of their intended action for the prisoner's dilemma. 69% of finalists who, before the show, state either to cooperate or to defect, actually choose their action consistently with their statement.
Within the pre-play, contestants strategically select their partner for the prisoner's dilemma. On the one hand, contestants are concerned about maximizing stake size for the prisoner's dilemma game, on the other hand, they are concerned about surviving the pre-play. Driven by the fear of being eliminated from the game, contestants may lie or bluff about stakes at play. But lies are revealed. The analysis shows that lying is punished by a higher propensity to be eliminated from the game.

Again, this behavior can be interpreted in terms of reciprocity. Lying is perceived as unkind behavior, and agents reciprocate unkindness with unkindness in form of voting liars off the game. In addition, contestants are bounded rational in the decision to lie in the sense that they essentially make the decision to lie contingent on the own position in the game, but the strengths and weaknesses of the other players are mostly neglected. Our findings highlight the relevance and informational value of communication, and enrich the literature with findings on non-verbal communication, pre-commitment, and modes of lying.

The third chapter continuous the investigation of cooperative behavior, but the center of attention is a special physical characteristic of people: their beauty.
There is empirical evidence that beautiful people are rewarded with significant economic benefits in form of higher salaries or better career prospects. The literature argues that so called "beauty-is-good" stereotyping can explain this "beauty premium." That is, people believe that physically attractive people are, for instance, more productive, more talented, or more trustworthy than less attractive people, and this stereotype-belief then mediates people's behavior towards the attractive.
In this chapter the effect of physical attractiveness on cooperative behavior is analyzed. Data from 211 episodes of the television game show "Golden Balls" are combined with data from independent facial appearance ratings of the show's contestants. In a survey, 728 people rated the facial appearance of 844 contestants on the basis of portrait photographs. These ratings are used to construct various measures of facial attractiveness for each contestant.
The results show that facially attractive contestants provoke cooperative behavior from their counterparts. This preferential treatment or beauty premium rewards attractive contestants with substantially higher monetary gains of up to £2 153, and the probability to obtain positive earnings increases by 5.9 percentage points. This finding applies across sexes, and is independent of other demographic characteristics, stake size, communication and past behavior. Further, the result is not driven by non-cooperative behavior of the attractive. With minor qualifications for younger and female contestants, facially attractive contestants are no more or less likely to cooperate than less attractive contestants.
However, the analysis reveals that the attractiveness effect vanishes in interactions between two contestants of the same sex. Contestants are more cooperative towards the attractive opponent only if the opponent is of the opposite sex. This suggests that stereotype beliefs about attractive people cannot explain our results consistently. Also, attractiveness has no effect in group-decisions made by the contestants prior to the prisoner's dilemma. People seem to have a preference to cooperate more with someone towards whom they are personally attracted. This preference reaches full effect when people interact in pairs and lack other relevant information about the opponent for taking their decision.

Chapter 1

Nash Implementation with Reciprocity Preferences

1 Introduction

The theory of implementation studies the problem of a planner who wishes to implement a socially optimal outcome, depending on the preferences over a set of possible outcomes of a number of agents. The implementation problem arises from the asymmetric information between the planner and the agents. In particular, the planner does not know the agents' true preferences over possible outcomes, and therefore must rely on the agents' individual reports in form of messages. These messages contain information about the agents' preferences. However, an agents' message can be a strategic lie in order to influence the outcome to her advantage. Hence, the planner designs a mechanism, a game-form, through which the agents' interact. Further, the planner defines a social choice function, which assigns an outcome to each possible profile of the agents' preferences. The solution to the implementation problem is then provided by a mechanism, such that every equilibrium has the outcome associated with the social choice function in the true state (for a survey of the literature see e.g., Jackson, 2001).
The theory usually assumes that agents have selfish preferences about outcomes. However, a large body of evidence from experimental and empirical research shows that people very often care about whether their behavior is perceived to be fair by one another, and therefore take into account underlying intentions behind their actions. A prominent idea is that of reciprocity, which describes peoples' willingness to sacrifice own material well-being in order to either punish or reward behavior by others that they perceive as unkind or kind, respectively (for a survey see e.g., Sobel, 2005).

In this study I take up the issue of reciprocity and incorporate a model of fairness and reciprocity into implementation theory. In particular, I investigate the implications of intention-based reciprocity preferences for Nash implementation (see Maskin, 1999). In modeling reciprocity I adopt the approach by Rabin (1993), who introduces the solution concept of *fairness equilibrium* in games of complete information. Agents are fairness utility maximizers, with the fairness utility specified as a function of the material payoff and a psychological payoff. This psychological payoff captures reciprocity motives, in the sense that an agent receives a positive psychological payoff from treating an other agent kindly (unkindly) when she believes that this agent is kind (unkind) towards her as well; otherwise the psychological payoff is negative. Hence, in a fairness equilibrium agents optimally respond to the actions and beliefs (about actions and beliefs) of the other agents, taking into account their own and the others' intentions. Further, the psychological payoff depends on (exogenously) given weights, that measure an agent's individual sensitivity for reciprocity, i.e., how strong kindness sensations affect the agent's behavior. These weights will have a decisive role in the later analysis.
With this solution concept at hand, I define a social choice function to be *Nash fairness implementable* if every *Nash fairness equilibrium* outcome of the game induced by a mechanism coincides with the allocation specified by the social choice function. Further, in line with Bierbrauer and Netzer (2012), I call a strategy profile a *psychologically robust equilibrium* if it is a Nash equilibrium irrespective of the agents willingness to trade-off material payoffs and reciprocal kindness, such that the agents' sensitivity for reciprocity does *not* affect (Nash) equilibrium behavior. Accordingly, a social choice function is *psychologically robustly implementable* if every psychologically robust equilibrium has the desired outcome, i.e., the one specified by the social choice function, and there exists no (other) Nash fairness equilibrium that has a different outcome.
In the first part of the analysis I focus on the canonical mechanism as introduced by e.g., Repullo (1987)), and investigate whether this mechanism allows for unique implementation in Nash fairness equilibrium. I consider three specific settings, which demonstrate the potential and limits of the applicability of the canonical mechanism, and establish the following results:

(1) in an environment in which the agents' preference orderings over outcomes form a Condorcet cycle, a Nash implementable social choice function (which assigns the preferred outcome to exactly one agent) is Nash fairness implementable provided the agents concern for reciprocity are small, i.e., below a given threshold. If the agents' concerns for reciprocity exceed that threshold, the canonical mechanism has no equilibrium. The intuition behind the result is the following: by the Condorcet cyclical ordering of preferences, in each state every agent top-ranks a different alternative, such that in equilibrium only one agent receives her preferred outcome. The other agents, who

do not receive their preferred outcome, have now strong motives to punish the one who gets the preferred outcome. This punishment bears costs in form of a lower material payoff, but also gains from a positive psychological payoff. Hence, for large values of reciprocity, the agents have an incentive to deviate from truth-telling, which, in turn, implies that there will be no truthful Nash fairness equilibrium.

(2) in an environment in which the agents have aligned preferences, a Nash implementable social choice function is not Nash fairness implementable in the canonical mechanism provided the agents' concern for reciprocity is large. In contrast to the result above, here reciprocity preferences raise the problem of multiple equilibria. In particular, the outcome specified by the social choice function is the one that all agents prefer the most in the true state, and the canonical mechanism yields this outcome whatever the agents' concern for reciprocity. However, if all agents coordinate on the (same) least preferred outcome, they are as unkind as possible towards each other. The mutual unkindness translates into a positive psychological payoff. It follows that the message profile with maximal unkindness constitutes a Nash fairness equilibrium for high values of reciprocity; and this, in turn, implies that the outcome of the canonical mechanism is not unique.

(3) finally, in an environment in which the agents' face a minority conflict in preferences, a Nash implementable social choice function is psychologically robustly implementable as the unique outcome of the canonical mechanism. Here, all but one agent top-rank the same alternative, and this alternative is the one that is associated with the social choice function in the true state. For the remaining agent this outcome is the least preferred one in the true state. The canonical mechanism then allows unique implementation, irrespective of the agents' concern for reciprocity.

Further, I derive a property that together with Nash implementability provides a sufficient condition for a social choice function to be *not* psychologically robustly implementable in the canonical mechanism, i.e., the canonical mechanism does not yield the desired outcome for higher values of reciprocity.
In the second part of this study I address the psychological robustness for general mechanisms. I establish that a Nash equilibrium of any mechanism that prohibits a single agent to affect the equilibrium outcome, is a psychologically robust equilibrium. However, those mechanisms cannot handle the problem of multiple equilibria. As soon as the possibility to affect the Nash equilibrium outcome is shut down, one obtains psychological robustness, but also non-uniqueness.

The rest of the study is organized as follows. In section 2 the related literature is discussed. Section 3 describes the setup, containing a formal mechanism design

framework, and defines the solution concepts of Nash fairness equilibrium and psychological robust equilibrium. Section 4 reviews Maskin's famous theorem for Nash implementation. Section 5 deals with the analysis of Nash fairness implementation in the canonical mechanism. Section 6 explores psychologically robust mechanisms. Finally, Section 7 concludes.

2 Literature

This study links two broad streams in the literature, (i) the theory of Nash implementation, and (ii) the behavioral approach of fairness and reciprocity.

(i) Nash implementation theory

The vast majority of the implementation literature follows mainstream economic theory by assuming that agents only care about (material) outcomes. In complete information environments, when agents are rational in their behavior, the most prominent idea is that of Nash equilibrium: each agent's action is an optimal response to the (given) actions of the other agents.
Maskin (1999, in circulation since 1977) first provided a solution to the implementation problem when there is complete information and the solution concept is Nash equilibrium. He proposes two conditions on the social choice function to be Nash implementable, namely monotonicity and no-veto power. Monotonicity is necessary for Nash implementation, and monotonicity coupled with no-veto power is sufficient for Nash implementability when there are at least three agents. A complete proof of Maskin's theorem was established by Wiliams (1986), Repullo (1987), and Saijo (1988), who introduce a canonical mechanism that allows for Nash implementation of any social choice function in any (possible) environment. The theory can be extended to multi-valued social choice functions, for which Maskin's theorem still applies. Moore and Repullo (1990) and Dutta and Sen (1991) provide a full characterization for social choice correspondences to be Nash implementable, which includes the two-person implementation problem. Additionally, several authors motivated replacements for the no-veto power condition, see e.g., Moore and Repullo (1990), Sjöström (1991), Danilov (1992).[1]

Several authors criticized the classic assumption of unbounded rationality of the agents, and investigated the implementation problem allowing for variations in the agents' behavior. Note that in the present work the agents do not depart from in-

[1]Besides, there has been considerable interest in implementation with more sophisticated equilibrium concepts, that require weaker conditions for implementation, see e.g., Moore and Repullo (1988). For a survey of the literature see e.g., Jackson (2001).

dividual rationality. It is assumed that agents have reciprocity preferences, which transform into rational strategies that are consistent with the rationality criteria of standard game-theory.

Learning One strand of literature is concerned with the issue of how the equilibrium of a mechanism is reached, and whether it is stable. Cabrales (1999) studies the problem of Nash implementation with "naive adaptive dynamics", in the sense that agents play the game repeatedly, whereby they can use the history from play to improve their strategies, which are not necessarily best responses. It turns out that agents adjust their strategies in the direction of better responses within a finite canonical mechanism, and the dynamics converge to a (stable) Nash equilibrium.[2] A theory of bounded rationality in mechanisms, where bounded rationality is modeled as the myopic behavior of agents to adjust their strategies to better responses, is established by Cabrales and Serano (2011). They provide a complete characterization for implementation in better-response dynamics.

Allowing for mistakes Sjöström (1993) considers trembling-hand perfect implementation, which offers an approach to modeling mistakes in implementation theory. A trembling-hand perfect equilibrium accounts for off-the-equilibrium path actions of the agents. The agents may choose (with negligible probability) unintended actions, which can be interpreted as modeling mistakes by the agents. Eliaz (2002) studies the implementation problem when some "faulty" agents make mistakes in the sense that they fail to choose equilibrium strategies correctly. The solution concept is a k-fault tolerant Nash equilibrium, which requires robustness to deviations from equilibrium by the faulty agents. Eliaz's approach is extended to incomplete information environments by Doghmi and Ziad (2009), who accordingly introduce the notion of k-fault tolerant Bayesian equilibrium.

Preference for honesty Further studies analyze mechanism design problem assuming that agents have process-regarding preferences, i.e., they care about the process of how outcomes are achieved. For instance, these agents have an intrinsic preference for honesty, and suffer a (small) utility loss if they lie in pursuit of wealth. In this spirit, Matushima (2008a) and Matushima (2008b) show that every social choice function is implementable in (Bayesian) Nash equilibrium, and that a small preference for honesty is sufficient to break down unwanted equilibria. Dutta and Sen (2011) and Lombardi and Yoshilhara (2011) provide a general characterization of Nash implementation with partially honest agents, i.e., at least one agent has a strict preference for revealing the true state over lying when truth-telling does not lead to a worse outcome than that which obtains when lying.

[2]In related work, Cabrales et al. (2003) tests the practical feasibility of a modified canonical mechanism in a laboratory experiment. When the strategy space is finite, and with the option to impose a fine on a dissident, the canonical mechanism successfully implements the social choice function in 68-80% of cases.

Related work is provided by Ben-Porath and Lipman (2012), and Kartik and Tercieux (2011), who study Nash implementation when the agents can provide some evidence about the true state of the world, such that the planner is able to discriminate between states.

(ii) Fairness and reciprocity

An ever-increasing theoretical and empirical literature emphasizes the limited ability of conventional game theoretic models to rationalize observed behavior, in which considerations about fairness and reciprocity play a role. Several theories try to explain reciprocal behavior assuming that utility functions are more complex in the sense that players care about more than just their own material payoffs. These theories can be classified into two categories: intention-based social preferences and outcome-based social preferences. In this study I focus on intention-based social preferences, i.e., preferences for reciprocity.[3]

Intention-based social preferences Reciprocity occurs when individuals act kindly as a reward for kind behavior by others, and act unkindly as a punishment for unkind behavior by others. Thereby, people view an action as kind or unkind by underlying intentions in addition to the consequences of an action. The formalization of intention-based reciprocity is pioneered by Rabin (1993) for static two-player normal form games. Kindness is evaluated at a norm, the equitable payoff, and a player cares about the opponent's material payoff only as a response to intentions. This requires as translation from material payoffs to fairness utility payoffs, in which players, in addition to material payoffs, receive a positive payoff from repaying kindness or punishing unkindness. Rabin's solution concept is a fairness equilibrium, in which strategies form best responses, and beliefs and actions are mutually consistent. That is, players hold beliefs about other players intentions as well as beliefs about other players possible alternatives. Since expectations matter, the appropriate framework is the one of psychological game theory. First defined by Geanakoplos et al. (1989), and later also by Battigalli and Dufwenberg (2009), in a psychological game payoffs dependent on both, actions and higher-order beliefs about actions. Segal and Sobel (2007) generalize Rabin's model providing an axiomatic foundation. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) extend Rabin's approach of reciprocal kindness to multi-player extensive form games with complete information. Their sequential reciprocity equilibrium requires that beliefs about intentions are sequentially revised correctly as the game proceeds. Levine (1998), Charness and Rabin (2002), and Cox et al. (2007) propose further modifications to Rabin's

[3]Theories that model outcome-based social preferences assume that people intrinsically care about their own and relative payoffs. Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) incorporate outcome-based social preferences in their models, where the utility is specified as a function of the players material payoff and the distribution of payoffs, such that fairness is modeled as self-centered "inequality aversion".

approach, incorporating both reciprocity and distributional concerns in one model. For a survey of the literature see Sobel (2005).

Intention-based Bayesian mechanism design More recently, Bierbrauer and Netzer (2012) study Bayesian mechanism design in the presence of intention-based social preferences. In a theoretical framework they define the solution concept of Bayes-Nash fairness equilibrium (BNFE), and model intention-based reciprocity following Rabin (1993), and Dufwenberg and Kirchsteiger (2004). In the first part of the paper, it is assumed that the social planner has information about how agents perceive and balance kindness against material payoffs, which is referred to as the exploration of "mechanisms with known kindness generating process". A pseudo revelation principle, that is a modified version of the fundamental revelation principle, is proven. It can be shown that under a non-direct mechanism, every efficient social choice function can be implemented in BNFE. In the second part of the paper, it is assumed that the social planner has no information about the weight agent's place on reciprocity. Attention is drawn to the psychological robustness of a mechanism with respect to reciprocal behavior of the agents. A mechanism is called psychologically robust, if it implements a social choice function in a Bayes-Nash equilibrium irrespective of any concerns for reciprocity by the agents. Bierbrauer and Netzer (2012) identify a sufficient condition, called "insurance property", for psychologically robust implementabililty of an incentive compatible social choice function. Further, they prove that for any Bayes Nash implementable social choice function, there exists an equivalent social choice function satisfying the insurance property; and, under suitable budget balance assumptions, there exists a psychologically robust mechanism.

I add to Bierbrauer and Netzer's (2012) approach, investigating the robustness of Nash implementation pioneered by Maskin (1999) with respect to intention-based reciprocity.

3 Setup

In the setup I largely follow the standard notation of Mas-Colell, Winston, and Green (1995), Chap. 23.

3.1 The social choice problem

Let $\langle I, X, \Theta \rangle$ represent the environment in which the (social) planner operates. There are a finite number of N agents indexed by $i \in I = \{1, ..., N\}$, that need to make a choice from some finite set of feasible alternatives X. Each agent

has a certain type $\theta_i \in \Theta_i$. A state of the world consists of a profile of types $\theta = (\theta_1, ..., \theta_N) \in \Theta$, with Θ the finite set of states, which defines a preference profile $\succeq^\theta = (\succeq_1^\theta, ..., \succeq_N^\theta) \in \mathcal{R}$, with $\mathcal{R}$ the set of possible profiles of preference orderings, on the set of feasible alternatives X. For instance, $x \succeq_i^\theta x'$ ($x \succ^\theta x'$) indicates that agent i weakly (strictly) prefers alternative x to x' in state θ.[4]
The environment is of complete information, i.e., the entire vector $\theta = (\theta_1, ..., \theta_N)$ is common knowledge among all agents, but not verifiable to the planner.
There is a material payoff function $u_i : X \times \Theta \mapsto \mathbb{R}$ representing agent i's preferences over alternatives in each particular state $\theta \in \Theta$.[5]
The planner's overall goal is defined by a social choice function (SCF) f, which determines a social desirable outcome $x \in X$ for each state $\theta \in \Theta$.

Definition 1 (Social Choice Function (SCF)). *A SCF $f : \Theta \mapsto X$ specifies an outcome $f(\theta) = x \in X$ for each state $\theta \in \Theta$.*

3.2 Mechanisms

In order to reach the desired allocation despite the agent's self-interest, the planner commits to a mechanism, that defines the available strategies and the method used to select an outcome based on the agents' strategies.

Definition 2. *A mechanism $\Phi = (M, g)$ consists of a message space $M = M_1 \times ... \times M_N$, and a mapping $g : M \mapsto X$.*

The mapping g is called the outcome function and specifies an alternative for every profile of messages.

3.3 Solution concepts

Fix a mechanism $\Phi = (M, g)$ and a profile of preferences $\theta \in \Theta$. For convenience, suppress the dependence on θ. In an environment of complete information, a mechanism $\Phi = (M, g)$ induces a normal-form game $\Gamma := [I, M, (\pi_i)_{i \in I}]$, with $m = (m_1, ..., m_N) \in M$ a message (strategy) profile[6], and $\pi_i : M \to \mathbb{R}$ the material payoff function for agent i.[7]
Let $B_{ij} = M_j$ denote the set of possible beliefs of agent i about agent j's action,

[4]For simplification I assume that the set of alternatives is the same in all states.

[5]The utility function induces a binary relation $\succeq^\theta$ on X, i.e., $x \succeq_i^\theta x'$ if and only if $u_i(x, \theta) \geq u_i(x', \theta)$.

[6]A message profile can also be written as (m_i, m_{-i}), where $m_i \in M_i$ denotes agent i's message and $m_{-i} \in M_{-i}$ is a list of messages sent by all other agents $j \neq i$. For expositional simplicity, I restrict attention to pure-strategies.

[7]Note that for any state $\theta \in \Theta$ a mechanism Φ induces a normal-form game Γ^θ, with $\Gamma^\theta \neq \Gamma^{\theta'}$ for all $\theta' \in \Theta$.

with $b_{ij} \in B_{ij}$ denoting agent i's first-order belief about the action of agent j.
Then agent i's *material payoff* from sending m_i, given her belief about the messages sent by all other agents, is equal to $\pi_i(m_i, (b_{ij})_{j\neq i}) = u_i(g(m_i, (b_{ij})_{j\neq i}), \theta)$.

(1) Nash equilibrium

The Nash equilibrium solution concept is by large the most well-known in game theory, given traditional assumptions about agents' preferences, rationality, and information available to the agents about each other.

Definition 3 (Nash Equilibrium (NE)). *Fix a state $\theta \in \Theta$. A message profile m^* of a mechanism Φ is a NE if, for every agent $i \in I$ it holds that*

$$m_i^* \in \arg\max_{m_i \in M_i} \pi_i(m_i, (b_{ij})_{j\neq i}), \quad \textit{with} \quad b_{ij} = m_j^* \quad \forall j \neq i.$$

A NE is a profile of strategies $(m_i^*)_{i\in I}$ such that for any agent $i \in I$ the action m_i^* is a best response to the actions that she believes the other agents will take; and in addition, these beliefs are correct.

(2) Nash fairness equilibrium

The Nash fairness equilibrium solution concept assumes that agents take into account the consequences and intentions of their actions. Agents are assumed to be fairness utility maximizer, i.e., in addition to material payoffs, they receive a positive (weighted) psychological payoff from being kind (unkind) to those, whose actions are perceived kind (unkind), and they derive a negative psychological payoff from being unkind (kind) to those whose actions are perceived kind (unkind). Consistently, reciprocal motivations depend directly on beliefs (about beliefs) in addition to which actions are chosen.
In the following I present Rabin's (1993) approach and apply his notion of Nash fairness equilibrium to mechanisms.
Let $C_{ijk} = B_{jk} = M_k$ denote the set of second-order beliefs, with $c_{ijk} \in C_{ijk}$ being agent i's belief about agent j's beliefs about agent k's behavior.
The kindness of the intention is evaluated at the *equitable payoff*. I specify the equitable payoff as the average between the largest and the smallest material payoff that agent i can potentially give to agent j by choice of her own strategy m_i, fixing the strategies of all agents $j \neq i$, and excluding Pareto inefficient strategies. Let $\pi_j^{e_i}((b_{ij})_{j\neq i})$ denote the equitable payoff for agent j from the perspective of agent i, given agent i's beliefs about the actions of the other agents, $(b_{ij})_{j\neq i}$. Formally,

$$\pi_j^{e_i}((b_{ij})_{j\neq i}) = \frac{1}{2}\left[\max_{m_i\in M_i} \pi_j(m_i, (b_{ij})_{j\neq i}) + \min_{m_i \in E_{ij}((b_{ij})_{j\neq i})} \pi_j(m_i, (b_{ij})_{j\neq i})\right],$$

where $E_{ij}((b_{ij})_{j\neq i}) \subseteq M_i$ denotes the set of conditionally and bilaterally Pareto efficient strategies.[8] Then, agent i's *kindness* towards agent j, depending on both, her own strategy m_i and her belief about all other agents' strategies $(m_{ij})_{j\neq i}$, is

$$\kappa_{ij}(m_i, (b_{ij})_{j\neq i}) = \pi_j(m_i, (b_{ij})_{j\neq i}) - \pi_j^{e_i}((b_{ij})_{j\neq i}),$$

Whenever the material payoff, π_j, that agent i is offering to agent j, exceeds the equitable payoff, $\pi_j^{e_i}$, then agent i thinks she is kind to agent j.
Similarly one can evaluate agent i's *perceived kindness*, that is, whether agent i believes that agent j is kind to her,

$$\lambda_{iji}(b_{ij}, (c_{ijk})_{k\neq j}) = \pi_i(b_{ij}, (c_{ijk})_{k\neq j}) - \pi_i^{e_j}((c_{ijk})_{k\neq j}),$$

with $\pi_i^{e_j}$ being the equitable payoff which agent i believes that agent j uses to evaluate her fairness towards herself. Note that $\lambda_{iji}(b_{ij}, (c_{ijk})_{k\neq j}) \equiv \kappa_{ji}(b_{ij}, (c_{ijk})_{k\neq j})$.
Now agent i's *fairness utility function* can be formally defined as follows,

$$U_i(m_i, (b_{ij}, (c_{ijk})_{k\neq j})_{j\neq i}) = \pi_i(m_i, (b_{ij})_{j\neq i}) + \sum_{j\neq i} \xi_{ij} \cdot \kappa_{ij}(m_i, (b_{ij})_{j\neq i}) \cdot \lambda_{iji}(b_{ij}, (c_{ijk})_{k\neq j}),$$

with $\xi_{ij} \geq 0$ an exogenously given weight measuring the extent of agent i's reciprocity concerns towards agent j. The fairness utility is a function of the agent's material payoff (first term) and a psychological payoff (left term), which is the sum of the product of agent i's own kindness and i's perceived kindness across all agents j. The multiplication of an agent's kindness and the expected kindness captures reciprocity motives – agent i's utility increases when treating agent j kindly (unkindly), $\kappa_{ij} > 0$ ($\kappa_{ij} < 0$), if she believes that agent j is kind (unkind) to her as well, $\lambda_{iji} > 0$, ($\lambda_{iji} < 0$)); by contrast, if agent i thinks that agent j is treating her kindly, $\lambda_{iji} > 0$, her utility is decreasing if she is unkind herself, $\kappa_{ij} < 0$, and vice versa.
Finally, in a Nash fairness equilibrium of the mechanism Φ, agents maximize fairness

[8] In the minimization part in the equitable payoff strategies are restricted to the subset of bilaterally Pareto efficient strategies, where efficiency is defined as a belief-dependent property. A strategy is called bilaterally Pareto inefficient if there exists another strategy m_i', which, conditional on the belief about the other agents' strategies $(b_{ij})_{j\neq i}$, yields no lower material payoff for any of the two agents, and a strictly higher material payoff for at least one of them. The restriction to a bilateral efficiency concept (and not a population-wide or global one) matches best the bilateral concept of reciprocity, i.e., agent i only cares about the behavior (intentions) of agent j, but not about the behavior of agent j towards a third agent k, such that efficiency is invoked conditional on the other agents' strategies. See Bierbrauer and Netzer (2012) for a discussion about the belief-dependent and bilaterally efficiency concept in line with Rabin (1993) compared to the belief-independent and globally efficiency concept in line with Dufwenberg and Kirchsteiger (2004). Also, in the specification of the equitable payoff one might not only assume a different efficiency concept, generalizations might be to allow for arbitrary weights $\alpha \in [0, 1]$ placed on the minimum and maximum material payoff, with the exogenous parameter α describing an agent's degree of satisfaction with the behavior of the others.

utilities, and strategies match beliefs correctly.

Definition 4 (Nash Fairness Equilibrium (NFE)). *Fix a state $\theta \in \Theta$. A message profile $m^* \in M$ of a mechanism Φ is a NFE if, for all agents $i \in I$, it holds that:*

$$m_i^* \in \arg\max_{m_i \in M} U_i(m_i, (b_{ij}, (c_{ijk})_{k \neq j})_{j \neq i}),$$

with $b_{ij} = m_j^*$, and $c_{ijk} = b_{jk} = m_k^*$, $\forall\ k \neq j$, $\forall\ j \neq i$.

In a NFE for any agent $i \in I$ the action m_i^* is the optimal choice, given correct first- and second-order beliefs about the action of the other agents.

(3) Psychologically robust equilibrium

Throughout it is assumed that the intensity of reciprocal preferences is common knowledge among agents, but not to the planner. The planner cannot measure the intensity of how an agent's kind (unkind) treatment raises (lowers) the weight that is placed on the other's material payoff, making an agent more (less) willing to sacrifice her own material payoff in favor (expense) of her opponent. Therefore, it is interesting to analyze the psychological robustness of an equilibrium, where psychological robustness is defined with respect to the agents' individual weights they place on reciprocity concerns, $\xi = (\xi_{ij})_{i,j \in I, i \neq j}$.
I adopt the approach of Bierbrauer and Netzer (2012) who introduce the notion of psychologically robust equilibria in a Bayesian setting. Analogously, I say that a profile of messages is a psychologically robust equilibrium, if it is a NE irrespective of the degree of reciprocity concerns that the agents might have, such that the agent's sensitivity for reciprocity does not affect equilibrium behavior.

Definition 5 (Psychologically Robust Equilibrium (PRE)). *Fix a state $\theta \in \Theta$. A message profile $m^* \in M$ of a mechanism Φ is a PRE if, for all agents $i \in I$, it holds that m^* is a NFE, for every $\xi \in [0, \infty)^{N(N-1)}$.*

Note that PRE is a refinement of NE. If $\xi_{ij} = 0$ for all agents $i, j \in I$ and $i \neq j$, then the agents act as selfish payoff maximizers and NFE coincides with NE.

3.4 Implementation

Since the planner does not observe the state of the world, her goal is to design a mechanism that induces the agents to truthfully reveal their preferences over outcomes. The implementation problem is solved when the set of equilibrium outcomes of the mechanism coincides with the allocation in each possible state of the world. Let Γ^θ denote the normal-form game induced by Φ in state $\theta \in \Theta$, and let $m^*(\theta)$ be a collection of NFE of the games Γ^θ, for all $\theta \in \Theta$.

Definition 6 (Nash fairness implementation (NFI)). *An SCF f is Nash fairness implementable if there exists a mechanism $\Phi = (M, g)$ such that, for any $\theta \in \Theta$ and for every NFE $m^*(\theta)$ of the normal-form game Γ^θ induced by Φ, it holds that: $f(\theta) = g(m^*(\theta))$.*

Hence, a social choice function is Nash fairness implementable if every Nash equilibrium yields the desired outcome.

Further, a social choice function is psychologically robustly implementable, if there is at least one PRE which has the desired outcome, and there is no NFE which has a different outcome.

Definition 7 (Psychologically robust implementation). *A SCF f is psychologically robustly implementable if there exists a mechanism $\Phi = (M, g)$ such that, for any $\theta \in \Theta$, there exists a PRE $m^*(\theta)$ of the normal-form game Γ^θ induced by Φ, with $f(\theta) = g(m^*(\theta))$, and there exists no NFE $\tilde{m}(\theta)$ with $g(\tilde{m}(\theta)) \neq g(m^*(\theta))$, for any $\xi \in [0, \infty)^{N(N-1)}$.*

4 Maskin's result

A challenge in implementation theory is to rule out undesired (untruthful) equilibria. The major breakthrough in characterizing uniqueness was the classic paper by Maskin (1999, in circulation since 1977), providing *necessary* and *sufficient* conditions for a SCF to be implementable in NE. The necessary condition for Nash implementation is monotonicity.

Definition 8 (Monotonicity). *An SCF f is monotonic if for any $\theta, \theta' \in \Theta$ and $x = f(\theta)$, but $x \neq f(\theta')$, for some $x \in X$, there exist some agent $i \in I$, and some $x' \in X$ such that $x \succeq_i^\theta x'$ and $x' \succeq_i^{\theta'} x$.*[9]

The definition follows the reasoning that, if an outcome $x \in X$ is the socially desired outcome in state θ, but not in state θ', then at least one agent must have reversed her preferences when moving from state θ to θ'.
Monotonicity together with a property called no-veto power is jointly sufficient for Nash implementability.

[9]The monotonicity requirement can be expressed in two (equivalent) ways. The second definition is as follows: a SCF f is monotonic if for any θ, θ' and $x = f(\theta)$ such that for each agent $i \in I$ and $x' \in X$ the relation $x \succeq_i^\theta x'$ implies $x \succeq_i^{\theta'} x'$ and $x = f(\theta')$. Intuitively, if an outcome is chosen by the SCF in state θ and moves up in each agent's preference ranking when moving to another state θ', then it should continue to be the desired alternative at θ'.

Definition 9 (No-veto power (NVP)). *An SCF f satisfies NVP if $x = f(\theta)$, whenever for some state θ and some alternative $x \in X$, we have $x \succeq_j^\theta x'$, for all $x' \in X$, and for all $j \in I$ except at most one.*

The no-veto power condition states that one agent alone cannot veto the majority view. If in some state θ an alternative x is top-ranked in the preference ordering by at least $(N-1)$ agents, then x should be chosen by the SCF.

Theorem 1 (Maskin, 1999). *(1) If a SCF f is implementable in NE, then it must be monotonic. (2) Given at least three agents ($N \geq 3$), any monotonic SCF f satisfying no-veto power is implementable in NE.*

The mechanism that has been frequently used to implement social choice functions in Nash equilibrium is the "canonical mechanism". Below the canonical mechanism is presented, following the design of Repullo (1987).

Canonical mechanism (Repullo, 1987) Denote the canonical mechanism by $\Phi^C = (M, g)$: each agent $i \in I$ simultaneously sends a message m_i, which consists of a state of the world $\theta \in \Theta$, an alternative $x \in X$, and an integer $z \in \mathbb{Z}^+$, with $\mathbb{Z}^+$ the set of non-negative integers, $m_i = (x^i, \theta^i, z^i) \in M_i$. With slight abuse of notation, a message is said to be the same for (at least) two agents if it coincides in the announced alternative and state, independent of the integer, i.e., I write $m_i = (x, \theta)$ whenever $m_i = m_j$ for some arbitrary integer z^i. The outcome function g of Φ is defined by the following three rules:

(1) If all N agents send the same message $m_1 = ... = m_N = (x, \theta)$, and $x = f(\theta)$, then the outcome is $g(m) = x$.

(2) If $(N-1)$ agents send the same message $m_i = (x, \theta)$, and $x = f(\theta)$, then the outcome is $g(m) = x$, except if the dissident j announces $m_j = (x^j, \theta^j, z^j)$ with $m_j \neq m_i$, then

   (i) $g(m) = x$ if $x^j \succ_j^\theta x$, and

   (ii) $g(m) = x^j$ otherwise.

(3) In all remaining cases, an integer game is played: the agent who announces the highest integer in her message wins in the sense that the outcome implemented is the alternative she proposes, i.e., $g(m) = x_i$, where i is such that $z^i \geq z^j$ for all $i, j \in I$, $i \neq j$. In case of ties, $g(m) = x_i$, where x_i is the alternative chosen by agent i, who is the agent with the lowest index among those agents who announced the highest integer.

The intuition of the canonical mechanism is as follows: if all agents reach a consensus such that the desired alternative is the one which the SCF associates with the true preference profile, then this alternative is implemented. If one agent disagrees and

proposes a different alternative, then there is a test that the alternative has to pass. If it passes the test, the alternative outcome is implemented, otherwise it is not. Often the agents whose messages pass the test are called *test agents* (see e.g., Moore, 1992; Cabrales, 1999). The outcome can only be changed if the test agent announces an alternative that she does not prefer to the one chosen by the other agents in the announced state. If more than one agent disagrees it cannot be inferred who is truthful and who is lying. The integer game then guarantees that there will always be one agent who has an incentive to deviate announcing a higher and higher integer, such that there will not exist an equilibrium.[10]

5 Implementation in Nash fairness equilibrium

The canonical mechanism allows implementation of social choice function that satisfy monotonicity and no-veto power as the unique Nash equilibrium outcome. In the following I ask to which extent a unique implementation in Nash fairness equilibrium is possible. I consider three specific settings, which show that the existence of a truthful Nash fairness equilibrium outcome as well as that the uniqueness of the equilibrium outcome is not guaranteed for higher values of reciprocity.
In particular, (i) when the agents' preference orderings form a Condorcet cycle in each state of the world, then the canonical mechanism allows a unique implementation in Nash fairness equilibrium for small values of reciprocity; otherwise there will be no Nash fairness equilibrium; (ii) when the agents' have aligned preferences over outcomes, then the canonical mechanism does not yield a unique outcome for high values of reciprocity, i.e., exceeding an upper-bound; and (iii) when the agents face a minority conflict in preferences, then a social choice function is indeed implementable in Nash fairness equilibrium irrespective of the agents' concern for reciprocity, furthermore, this social choice function is also psychologically robustly implementable.
In addition, for the Condorcet setting, a detailed analysis of the determinants that specify the "degree of robustness" of the truth-telling Nash fairness equilibrium outcome with respect to upper-bounds on the agents' weights for reciprocity is provided.

[10]Many of the (constructive) proofs in the implementation literature employ integer games or replace the integer game by a modulo game (see e.g., Saijo, 1988; Danilov, 1992) to eliminate unwanted equilibria. Integer games have no pure-strategy equilibria, since any agent can obtain her most preferred outcome by the choice of an higher integer. In the modulo game each agent announces an integer from a finite set and the agent whose index matches the sum of the integers modulo the number of agents receives the allocation she has announced. By the same logic as of the integer game, if the outcome is not best for each agent, then there will always be an agent who wishes to announce a different, i.e., a higher integer. However, the unnatural feature of both integer and modulo games is often criticized in the literature, see e.g. Moore (1992), and Jackson (1992).

In particular, I determine the effects of ceteris paribus changes of the parameters that specify the agents' expected loss and benefit from reciprocal behavior.
Finally, a property, which I call "conflict property", is derived, that together with Nash implementability provides a sufficient condition for a social choice function to be *not* psychologically robustly implementable. In other words, there exists no psychologically robust equilibrium which has the desired outcome in the true state.

Note, in the following the terms "desired outcome" or "truth-telling outcome" have similar meaning in the sense that they denote the outcome which the social choice function associates with the true state of the world. By the same reasoning, a "truth-telling message profile" denotes the profile in which all agents coordinate on the desired outcome.

5.1 Limitations of the canonical mechanism

Let's start with an abstract setting in which the agents' preferences over outcomes form a Condorcet cycle in each state of the world, i.e., in and within every state of the world each agent top-ranks a different alternative. Consider Example 1.

Example 1 (Condorcet cycle). *Suppose there are three agents, $i \in \{1,2,3\}$, a set of three alternatives $\{x_1, x_2, x_3\}$, and three possible states of the world, $\{\theta_1, \theta_2, \theta_3\}$. The agents' preference orderings over alternatives in the three states form a 3×3 Condorcet cycle described by the following matrix,*

states	θ_1			θ_2			θ_3		
agents	*1*	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
	x_1	x_2	x_3	x_3	x_1	x_2	x_2	x_3	x_1
alternatives	x_2	x_3	x_1	x_1	x_2	x_3	x_3	x_1	x_2
	x_3	x_1	x_2	x_2	x_3	x_1	x_1	x_2	x_3

such that, for instance, in state θ_1 agent 1 has the preferences ordering $x_1 \succ x_2 \succ x_3$. Due to the cyclical order of preferences among agents and states, in each state every agent prefers a different alternative the most *and the* least. *Hence, in each state any single-valued SCF must assign the best and the worst alternative in the agents' preference ordering to one of the agents. The SCF f is defined as follows:*

$$f(\theta_k) = x_k, \quad \textit{with} \quad k \in \{1,2,3\}.$$

The SCF satisfies Maskin's conditions of monotonicity and NVP and is therefore implementable in NE.[11]*. Suppose there exists a material equivalent for the utility enjoyed with each outcome. Let $\pi_i \in \{H, M, L\}$, denote the material payoff, where*

[11] The proofs are relegated to Appendix A.3.1

H denotes the highest possible material payoff, $M = \delta H + (1-\delta)L$, with $\delta \in (0,1)$, the medium payoff, and L the lowest possible payoff. Since agents' preference orderings and appropriate monetary equivalents cycle within the three states, preferences among agents are symmetric.

Obviously, in this Condorcet example there is always one agent who receives the lowest material payoff in the allocation prescribed by the social choice function, such that truth-telling might not be an obvious choice. In the Nash equilibrium the agents for whom the outcome does not yield the highest material payoff might interpret the behavior of the one agent who receives her top-ranked alternative as unkind. If punishment is possible, depending on the rules of the canonical mechanism, then agents may be willing to punish this unkind behavior, provided that reciprocity motives are strong enough. Indeed, it can be shown that truth-telling is a Nash fairness equilibrium outcome if and only if the agents' sensitivity for reciprocity is limited to an upper-bound; otherwise, there exists no equilibrium. The following proposition states the result.

Proposition 1. *Consider Example 1. The social choice function f is implementable in NFE in the canonical mechanism if and only if*

$$\xi_{ii'} \leq \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)} \equiv \varepsilon_1, \quad \forall i, i' \in I, i' \neq i.$$

Otherwise, there exists no NFE.

A formal proof of Proposition 1 can be found in Appendix A.3.3; for a detailed description of the rules of the canonical mechanism specific for the example I refer to Appendix A.1.
The logic of the result is as follows. Suppose all agents announce the truth-telling message profile m^*. Then there is always one agent who either receives the lowest (agent L), or the intermediate (agent M), or the highest (agent H) material payoff in the allocation prescribed by the social choice function; and for all pairs of agents, at least one of them is kindness-neutral towards the other. In particular, agent L is kindness-neutral towards both agent H and M, since she cannot change the outcome by the rules of Φ^C; and agent M is kindness-neutral towards agent H, since all outcomes she can induce by the rules of Φ^C yield Pareto inefficient outcomes and therefore equilibrium payoffs and equitable payoffs coincide. Hence, the fairness utilities from truth-telling equal the material payoffs, i.e, $U_i(m^*) = \pi_i(m^*)$, for all agents $i \in I$. However, by choosing m_H^* agent H gives agent M the lowest material payoff. Agent M now thinks that agent H is being unkind: agent H is a test agent and could have chosen a message, which had changed the outcome of Φ^C in favor for agent M. Therefore agent M is willing to be unkind as well. Dependent on the size of $\xi_{MH} > 0$, agent M may find it optimal to deviate from truth-telling in order to punish agent H for her unkind behavior. Although this punishment makes herself

materially worse off, agent M derives satisfaction from a positive psychological payoff. Therefore, the size of ξ_{MH} must be restricted to a certain cutoff value $\varepsilon_1 > 0$, such that m^* is a NFE, otherwise m_M^D maximizes $U_M(m_M, m_{-i}^*)$, which implies that truth-telling is not a NFE. Thus, the truth-telling message profile m^* is a NFE if and only if $\xi_{MH} \leq \varepsilon_1$. Because of the cyclical order of preferences, any other (untruthful) message profiles $m \neq m^*$ will have an outcome that gives one agent the highest, one agent the intermediate, and one agent the lowest material payoff (in the true state); but now agent L will be a test agent. Agent L always has an incentive to change the outcome to her advantage, and punishes the others for their unkind behavior. Thus, there exist no other NFE, for all $\xi_{ii'} \in [0, \infty)^6$.

Example 1 shows that, in a Condorcet setting, a social choice function is implementable as the outcome of a canonical mechanism, if and only if reciprocity preferences are not too strong. In particular, if the agents' motives for reciprocity are below an upper-bound, i.e., $\xi_{ii'} \leq \varepsilon_1$, for $i, i' \in I$, $i \neq i'$, then reciprocity-induced punishment becomes unattractive or impossible. But, if the agents motives for reciprocity are strong, i.e., $\xi_{ii'} > \varepsilon_1$, for $i, i' \in I$, $i \neq i'$, then there exists no equilibrium. Note that there is no problem with unwanted or multiple equilibrium outcomes. If reciprocity concerns are large enough, then the mechanism does not yield the desired equilibrium outcome and there is also no other equilibrium outcome irrespective of the weight $\xi_{ii'}$. Hence, if the upper-bound on the reciprocity weights, which depends on the exogenous parameters chosen for the material payoff, is very low, then already a small sensitivity for reciprocity is enough to eliminate any equilibria. For a detailed discussion of the psychological incentives in Example 1 see Section 5.3.

The next example demonstrates that the desired outcome does not need to be the unique outcome of the canonical mechanism, if the agents' concerns for reciprocity are large enough. In this setting the agents have aligned preferences over alternatives. In particular, in every state each agent has the same ranking over alternatives, and the outcome associated with the social choice function is the one which is preferred by all agents.

Example 2 (Aligned Preferences). *Consider an environment with three agents, $i \in \{1, 2, 3\}$, a set of three alternatives $\{x_1, x_2, x_3\}$, and two possible states of the world, θ_1, θ_2, such that in every state the agent's preferences among outcomes are aligned.*

states	θ_1			θ_2		
agents	*1*	*2*	*3*	*1*	*2*	*3*
	x_1	x_1	x_1	x_2	x_2	x_2
alternatives	x_3	x_3	x_3	x_3	x_3	x_3
	x_2	x_2	x_2	x_1	x_1	x_1

Suppose the SCF f is defined as follows:

$$f(\theta_k) = x_k, \quad \textit{with} \quad k \in \{1, 2\},$$

such that in each state the outcome chosen by f is the socially and collectively desired one. Trivially, the SCF f is monotonic and satisfies NVP, so that it is Nash implementable. Again, suppose there exists a material equivalent for the utility enjoyed with each outcome, denoted by $\pi_i \in \{H, M, L\}$, with $H > M > L$, and $M = \delta H + (1 - \delta)L$, with $\delta \in (0, 1)$.

In Example 2 the truthful NE outcome is clearly the desired one, i.e., the outcome that is associated with the social choice function in the true state. However, when the agents' concern for intentions is large, then there might exist other equilibria. Consider the profile in which all agents coordinate on the least preferred alternative in the true state. The outcome results in the lowest material payoff for each agents, and they are maximal unkind towards each other. The psychological payoff from mutual unkindness may now offset the costs associated with the low material payoff, provided reciprocity concerns are large enough. Indeed, the profile with "maximal unkindness" is a Nash fairness equilibrium of the canonical mechanism, iff the agents' concern for intentions are strong enough, and this equilibrium has an outcome which is not the desired one. The next proposition shows the result.

Proposition 2. *Consider Example 2 and the canonical mechanism. If ξ is such that*

$$\sum_{i' \neq i} \xi_{ii'} \geq \frac{1}{H - L} \equiv \varepsilon_2, \quad \forall i, i' \in I, i' \neq i,$$

then the social choice function f is not implementable in NFE in the canonical mechanism, due to non-uniqueness of the equilibrium outcome.

Again, a formal proof of Proposition 2 can be found in Appendix A.3.4. The intuition is the following. In each state all agents top-rank the same alternative, and this alternative is the one associated with the social choice function. When all agents coordinate on the desired outcome, then the canonical mechanism yields the desired outcome which gives each agent the maximum material payoff. Further, all pairs of agents are kindness-neutral towards each other, which implies that fairness utilities coincide with material payoffs, $U_i(m^*) = \pi_i(m^*)$. Since beliefs are all zero, no agent can profit from a unilateral deviation. Any unilateral deviation from truth-telling results in a lower material payoff than the one obtained from truth-telling. Hence, for every $\xi \in [0, \infty)^6$ the message profile m^* is a NFE.
Now, suppose the true state is θ_1, and that all agents agree on the same message $\tilde{m}_i = (x_2, \theta_2)$. By rule (1) of Φ^C, outcome x_2 is implemented, which yields for each agent the lowest material payoff (since the true state is θ_1). This makes all

pairs of agents as unkind as possible towards each other. However, being mutually unkind yields a positive psychological payoff for each agent. Consider now a possible unilateral deviation $m_i^D \neq \tilde{m}_i$ such that the outcome of Φ^C changes. Any such message m_i^D is now interpreted as kind. The outcome yields a higher material payoff than the one resulting from $\tilde{m}$, but, given beliefs, a negative psychological payoff. It can be shown that if the size of the sum of the reciprocity weights is large enough, $\sum_{i \neq i'} \xi_{ii'} \geq \varepsilon_2$ for all $i, i' \in I$, and $i' \neq i$, then the "benefit" from a larger material payoff cannot offset the loss from a negative psychological payoff. This, in turn, implies that the profile $\tilde{m}$ constitutes, in addition to m^*, a NFE of Φ^c, iff $\sum_{i \neq i'} \xi_{ii'} \geq \varepsilon_2$, with an outcome unequal to the desired outcome.

Example 2 demonstrates that the assumption of reciprocity preferences can raise the problem of multiple equilibria in the canonical mechanism. Although there exists a NFE that has the desired outcome for any $\xi \in [0, \infty)^{N(N-1)}$, the equilibrium outcome is not unique.

To sum up, Example 1 and Example 2 illustrate that Maskin's results about implementation have only limited applicability when the solution concept is NFE. When agents behave in a reciprocal manner, such that they are inclined to respond positively or negatively toward the actions of others, then the canonical mechanism does not necessarily yield the desired outcome. For strong reciprocity motives either truth-telling is not an equilibrium outcome, or truth-telling is not the unique equilibrium.

5.2 A possibility result

In this section a possibility result is derived. In the environment defined by the next example, the social choice function can indeed be implemented as the desired equilibrium outcome of the canonical mechanism, whatever the agents concerns for reciprocity. In the setting, agents face a "minority conflict" among preferences over alternatives. In particular, there are $(N-1)$ agents who all top-rank the same alternative in their preference orderings, and there is exactly one agent who least prefers this alternative. Consider Example 3.

Example 3 (Minority conflict). *Consider an environment with three agents, $i \in \{1, 2, 3\}$, a set of three alternatives $\{x_1, x_2, x_3\}$, and two possible states of the world, θ_1, θ_2. In every state there are two agents who (globally) top-rank the same alternative, whereas this alternative is (globally) lowest-ranked for a third agent.*

states		θ_1			θ_2	
agents	*1*	*2*	*3*	*1*	*2*	*3*
	x_1	x_1	x_2	x_2	x_1	x_2
alternatives	x_3	x_2	x_3	x_3	x_3	x_1
	x_2	x_3	x_1	x_1	x_2	x_3

Define the SCF f by

$$f(\theta_k) = x_k, \quad \textit{with} \quad k \in \{1, 2\},$$

such that the SCF is implementable in NE.[12] *Again, suppose there exists a material equivalent for the utility enjoyed with each outcome, denoted by $\pi_i \in \{H, M, L\}$, with $H > M > L$, and $M = \delta H + (1 - \delta)L$, but now $\delta \in (\frac{1}{2}, 1)$.*

The minority conflict among agents' preferences makes it possible that the outcome of the canonical mechanism will be the desired outcome, whatever the agents' willingness to trade-off material payoffs and kindness sensations. However, there will be more than one psychologically robust equilibrium, but all those equilibria have the desired outcome. Proposition 3 states the result.

Proposition 3. *Consider Example 3. The social choice function f is psychologically robustly implementable, i.e., the SCF is implementable as the unique NFE outcome of the canonical mechanism, for every $\xi \in [0, \infty)^{N(N-1)}$.*

For the formal proof of Proposition 3 I refer to Appendix A.3.5; a detailed description of the rules of the canonical mechanism specific for Example 3 can be found in Appendix A.2.
To see the logic behind Proposition 3 suppose all agents announce the truth-telling message m_i^*. Then the outcome results in the highest payoff for two agents i and i', and the lowest material payoff for a third agent i''. Since both agents i and i' receive the highest material payoff, any other choice would yield a lower material payoff for both of them, implying that they are kindness-neutral towards each other. Agent i'' cannot change the outcome by the rules of Φ^C, which implies that agent i'' is kindness-neutral towards both agent i and agent i'. Hence, for all pairs of agents at least one of them is kindness-neutral, such that $U_i(m^*) = \pi_i(m^*)$. Consider agent i and agent i' who can potentially change the outcome of the canonical mechanism. Since truth-telling delivers the maximum material payoff for both, and since the corresponding beliefs are zero, any deviation would yield a fairness utility below the one from truth-telling. Thus, there is no profitable deviation, implying that truth-telling is a PRE.
Now consider a profile of messages in which agent i and agent i' announce the truth,

[12] The SCF is Maskin monotonic and satisfies NVP, for the proofs see Appendix A.3.2.

but agent i'' does not, such that rule (2) applies. However, agent i'' cannot change the outcome by the rules of Φ^C, and the outcome associated with the social choice function in the true state will be implemented. Again, the outcome yields the highest material payoff for both agent i and i', and for all pairs of agents at least one of them is kindness-neutral. Hence, all profiles m with $m_{i''} \neq m_i^* = m_{i'}^*$, are also psychologically robust equilibria, and yield the same outcome, that is the one associated with the true message profile. It can be shown that any other message profile cannot be supported by NFE, for every $\xi \in [0, \infty)^6$.

Summarizing, it has been shown that in a setting with minority conflict among agents' preferences, the canonical mechanism is the appropriate tool for truthful implementation in NFE, whatever the agents' concern for reciprocity.

5.3 The incentives for reciprocity in Example 1

In the following, attention is drawn again to Example 1 of Section 5.1. I analyze in more detail the determinants of the degree of "robustness" of the truthful Nash fairness equilibrium with respect to the agents' reciprocity weights. Recall from Example 1 that the canonical mechanism yields an outcome that is top-ranked for exactly one agent. In equilibrium, the behavior of this agents is perceived as unkind by the other two agents. One of these two agents might then have the possibility to reciprocate the perceived unkindness, provided her weight placed on reciprocity exceeds an upper-bound. The upper-bound is determined by exogenous parameters, that describe the (relative) difference between agents' material gains resulting from the outcome of the canonical mechanism. Therefore, changes to those parameters have an impact on the strength by which kindness sensations affect behavior.
In the first part of this section, the costs and benefits from punishment are derived, as functions of the exogenous parameters δ, and $(H-L)$. In a second step, a comparative statics analysis is accomplished, considering ceteris paribus changes of those parameters. It can be shown that the incentives for punishment vary substantially.

5.3.1 Costs and Benefits of Punishment

Recall from the proof of Proposition 1 in Appendix A.3.3, when all agents coordinate on the truth-telling message profile, the mechanism yields the desired outcome, which results in the highest material payoff for agent H, the intermediate material payoff for agent M, and the lowest material payoff for agent L. In particular, $U_H(m^*) = \pi_H(m^*) = H$, $U_M(m^*) = \pi_3(m^*) = M = \delta H + (1-\delta)L$, and $U_L(m^*) = \pi_L(m^*) = L$, with $\delta \in (0,1)$, and $H > M > L$.
However, truth-telling by agent H is unkind towards both agent M and agent L. Let us focus on the unkindness towards agent M: agent H (hypothetically) could have changed the outcome in favor for agent M, i.e., to her top-ranked outcome, giving

agent M the highest material payoff. Hence, $\kappa_{HM}(m^*) = \frac{1}{2}(1-\delta)(L-H) < 0$.

- Unkindness of agent H towards agent M: $\mathcal{U} \equiv \frac{1}{2}(1-\delta)(L-H) < 0$.

Therefore, agent M has an incentive to punish agent H for being unkind, and, since agent M is a test agent, she can change the outcome of the canonical mechanism to be the intermediate one for agent H, i.e., $\pi_H(m_H^*, m_M^D, m_L^*) = \delta H + (1-\delta)L = M < H$, and $\kappa_{MH}(m_H^*, m_M^D, m_L^*) = \frac{1}{2}(1-\delta)(L-H) < 0$. Given beliefs, the punishment of agent M towards agent H can be determined as follows,

- Punishment of agent M towards agent H: $\mathcal{P} \equiv \frac{1}{2}(1-\delta)(L-H) < 0$.

However, the punishment goes along with costs and benefits for agent M. Compare the fairness utilities of agent M, i.e., $U_M(m^*) = M$ and $U_M(m_H^*, m_M^D, m_L^*) = L + \xi_{31}\frac{1}{4}(1-\delta)^2(H-L)^2$. On the one hand, agent M benefits from punishment, since punishing unkindness with unkindness results in an additional positive psychological payoff, $\xi_{MH}\frac{1}{4}(1-\delta)^2(H-L)^2$.

- Benefit from punishment for agent M: $\mathcal{B} \equiv \xi_{MH}\frac{1}{4}(1-\delta)^2(H-L)^2$.

On the other hand, agent M suffers a loss through a lower material payoff, i.e., $|M-L| = \delta(H-L)$.

- Loss from punishment for agent M: $\mathcal{L} \equiv \delta(H-L)$.

Hence, as soon as the benefits, $\mathcal{B}$, exceed the loss from punishment, $\mathcal{L}$, agent M does no longer stick to the truth-telling strategy, and the truth-telling NFE breaks down.

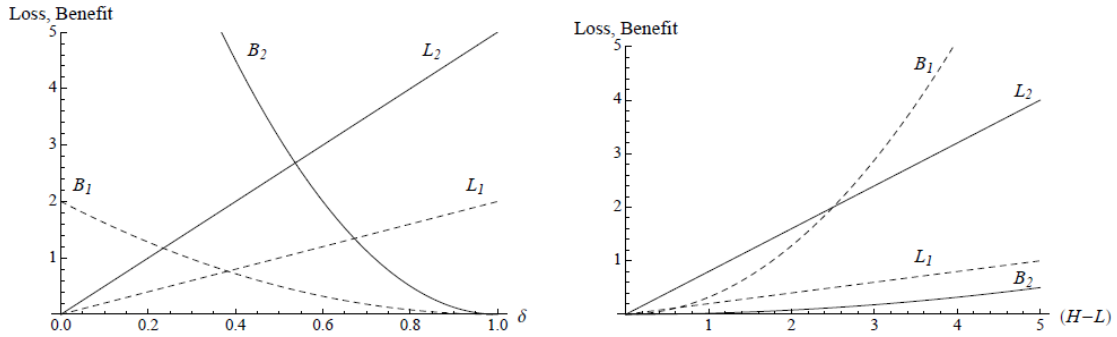
5.3.2 Comparative Statics

Consider ceteris paribus changes of the parameters $(H-L)$ δ, that determine agent M's loss, $\mathcal{L}$, and benefit, $\mathcal{B}$, from punishment. Figure I and Figure II illustrate the results.

Focus on $(H-L)$. The spread between the best and worst material payoff, $(H-L)$, reflects the intensity of the agents' preferences, and serves as an objective measure of absolute stakes size. Concerning ceteris paribus change with respect to $(H-L)$, keeping δ and ξ_{MH} fixed, we have $\frac{\partial \mathcal{L}}{\partial (H-L)} = \delta > 0$ and $\frac{\partial \mathcal{B}}{\partial (H-L)} = -\xi_{MH}\frac{1}{2}(1-\delta)^2(H-L) > 0$. That is, with an increase in $(H-L)$ both, the loss and benefit from punishment increase.

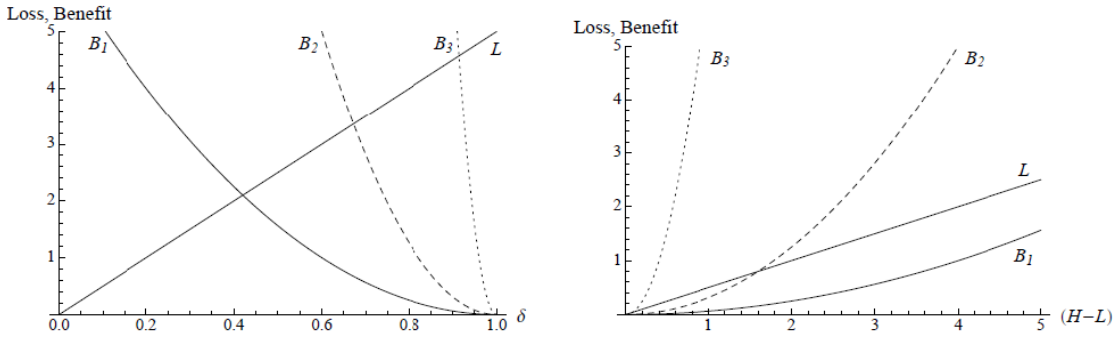
Focus on δ. The parameter $\delta \in (0,1)$ reflects the relative difference between the agents' intermediate material payoff and the highest and lowest material payoff. For large δ, the difference between the agents' highest and intermediate payoff is relatively small. In the extreme case, $\delta \to 1$, both payoffs almost equal each other, and the agents nearly become indifferent between the most preferred and intermediate preferred alternative. Concerning ceteris paribus changes with respect to δ, keeping

Figure I: Loss and benefit from punishment for fixed ξ_{MH}



Note: B and L describe the loss and benefit from punishment, respectively, keeping ξ_{MH} fixed. The left figure shows B and L as functions of δ for two different levels of $(H-L)$, i.e., B_1 and L_1 for $(H-L)=2$ (dashed lines), and B_2 and L_2 for $(H-L)=5$ (solid lines). For larger values of $(H-L)$, both B and L increase. The right figure shows B and L as functions of $(H-L)$, for two different levels of δ, i.e., B_1 and L_1 for $\delta=0.2$ (dashed lines), and B_2 and L_2 for $\delta=0.8$ (solid lines). For larger values of δ B decreases and L increases, such that for $\delta=0.8$ L lies above B.

Figure II: Loss and benefit for varying values of ξ_{MH}.



Note: B and L describe the loss and benefit from punishment, respectively, for varying levels of ξ_{MH}. The left figure shows B and L as functions of δ, keeping $(H-L)$ fixed. B_1 for levels of $\xi_{MH}=1$ (solid line), B_2 for levels of $\xi_{MH}=5$ (dashed line), and B_3 for levels of $\xi_{MH}=100$ (dotted line); L is independent of ξ_{MH}. The loss from punishment increases in δ and for higher levels of ξ_{MH}. The right figure shows B and L as functions of $(H-L)$, keeping δ fixed. B_1 for levels of $\xi_{MH}=1$ (solid line), B_2 for levels of $\xi_{MH}=5$ (dashed line), and B_3 for levels of $\xi_{MH}=100$ (dotted line); L is independent of ξ_{MH}. The larger ξ_{MH}, the larger the incentive for punishment. The benefit from punishment increases in $(H-L)$ and for higher levels of ξ_{MH}.

$(H-L)$ and ξ_{MH} fixed, we have $\frac{\partial \mathcal{L}}{\partial \delta} = H-L > 0$ and $\frac{\partial \mathcal{B}}{\partial \delta} = -\xi_{MH}\frac{1}{2}(H-L)^2(1-\delta) < 0$, i.e., with an increase in δ the loss from punishment increases, but the benefit from punishment decreases. Hence, if δ is sufficiently large, then the loss from punishment will exceed the benefit, which eliminates agent M's incentive to deviate from truth-telling.

Focus on ε_1. Finally, consider the effects of ceteris paribus changes on the agents' critical reciprocity weight $\varepsilon_1 = \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)}$, i.e.,

$$\frac{\partial \varepsilon_1}{\partial \delta} = \frac{4}{(H-L)}\left[\frac{1}{(1-\delta)^2} + \frac{2\delta}{(1-\delta)^3}\right] > 0,$$
$$\frac{\partial \varepsilon_1}{\partial (H-L)} = -\frac{4\delta}{(1-\delta)^2}\frac{1}{(H-L)^2} < 0.$$

An increase in δ results in an increase of ε_1. On the one hand, keeping $(H-L)$ fixed, and letting $\delta \to 1$ for $\delta \in (0,1)$, we have $\lim_{\delta \to 1} \varepsilon_1 = \infty$, such that in the limit case the agents' weight for reciprocity is nearly unbounded. This implies that agents can have a strong preference for reciprocity which does not destroy the truth-telling NFE. On the other hand, for $(H-L)$ fixed, and $\delta \to 0$ for $\delta \in (0,1)$, we have $\lim_{\delta \to 0} \epsilon = 0$. Here, even if the agents' have very small preferences for reciprocity, the NFE breaks down: If δ is near zero, then the upper-bound on the agents' weight for reciprocity is near zero, such that for already very small reciprocity preferences, truth-telling is no NFE.
An increase in $(H-L)$ results in a decrease of ε_1. The larger the spread between stake size of the best and worst material payoff, the more restrictive the upper bound on the agents' reciprocity weights. In reverse to the effects of δ above, it becomes beneficial for a particular agent to deviate from the truth-telling NFE when $(H-L)$ is relatively large.

5.4 On the existence of a truthful NFE

In this section I take up the issue of the existence of a truthful NFE in the canonical mechanism. As Example 1 demonstrates, there are settings in which no Nash fairness equilibrium exists for large values of reciprocity. Based on this observation, I define a general property, which is called *conflict property*, that together with Nash implementability provides a sufficient condition for a social choice function to be *not* implementable as the desired equilibrium outcome of the canonical mechanism. Note that this does not imply that there exists no NFE; it is just sufficient for truth-telling to be not an equilibrium outcome.

Consider a social choice function that assigns an outcome to any possible profile of the agent's preferences in a way that this outcome is "good" for agent i, but

"bad" for agent j. Then agent i and j have conflicting preferences over outcomes. If both, agent i and j, are also test agents, i.e., they can affect the other's payoff, then agent j might be willing to sacrifice own material payoff in order to punish agent i. Provided reciprocity motives are strong enough, this will destroy the truth-telling equilibrium outcome.
Formally, the conflict property is defined as follows.

Definition 10 (Conflict). *Given an environment $\langle I, X, \Theta \rangle$. A SCF f has the conflict property if there exist $\theta \in \Theta$, $i, j \in I, i \neq j$, and $x, x', x'' \in X$ such that*

(1) $x = f(\theta)$;

(2) $x \succ_i^\theta x'$ *and* $x \succ_i^\theta x''$;

(3) $x'' \succ_j^\theta x \succ_j^\theta x'$.

Statement (1) defines the SCF f. Statement (2) and (3) point out the conflict between agent i and j in state θ. Statement (2) specifies that agent i prefers the outcome associated with the social choice function to two other alternatives; statement (3) ensures that the outcome associated with the social choice function is neither top-ranked nor bottom-ranked for agent j. Hence, if all agents coordinate on the same state and outcome associated with the social choice function, then agent i and j are both test agents.

Further, in any NE bilateral kindness between agents cannot be positive. The lemma below shows the result, which was originally derived by Bierbrauer and Netzer (2012) in the case of Bayes Nash equilibrium.[13]

Lemma 1. *For any NE profile m^* it holds that $\kappa_{ij}(m^*) \leq 0 \; \forall \; i, j \in I, \; j \neq i$.*

The conflict property together with Lemma 1 implies the result, which is stated in the following proposition.

Proposition 4. *Consider an environment $\langle I, X, \Theta \rangle$. Any social choice function f that is implementable in NE and satisfies the conflict property is not psychologically robustly implementable as the desired outcome of the canonical mechanism Φ^C.*

Proof. Consider the canonical mechanism Φ^C. Suppose the true state is θ, such that m^* yields the NE outcome x of Φ. Then it holds that $\pi_i(m_i', m_{-i}^*) \leq \pi_i(m_i^*, m_{-i}^*)$ for any $m_i' \neq m_i^* \in M_i$ of agent i. From Lemma 1, it holds that $\kappa_{rs}(m^*) \leq 0$, for all agents $r, s \in I$, $r \neq s$. Lemma 1 together with the conflict property imply that there exists agent i and j such that $\kappa_{ij}(m^*) < 0$, i.e., in state θ, for agent i we have $x \succ_i^\theta x''$, but for agent j we have $x \prec_j^\theta x''$, such that agent i could have changed the

[13] For the proof see Lemma 5 in Appendix A of Bierbrauer and Netzer (2012), pp. 47-48, which can be applied to NE in a straightforward manner.

outcome of Φ^C in favor of agent j.
Consider a profile ξ, where $\xi_{ji} > 0$, $j \neq i$, and $\xi_{jk} = 0$, for all $k \neq j, i$. Condition (3) of the conflict property ensures that there exists a message $m_j'' \neq m_j^* \in M_j$ for agent j that yields outcome x'. This implies $\kappa_{ji}(m_j'', m_{-j}^*) < 0$, $i \neq j$, since $x' \prec_i^\theta x$. But then agent j has an incentive to deviate from the truth-telling message m_j^* to m_j'' iff

$$\xi_{ji} > \frac{\pi_j(m^*) - \pi_j(m_j'', m_{-j}^*)}{\kappa_{ij}(m^*)[\kappa_{ji}(m_j'', m_{-j}^*) - \kappa_{ji}(m^*)]} \equiv \bar{\varepsilon}.$$

Hence, there always exists a lower bound $\bar{\varepsilon} > 0$ on the weight ξ_{ji} such that for $\xi_{ji} > \bar{\varepsilon}$ the message profile m^* is not a NFE. □

The conflict property together with implementability in NE are jointly sufficient for non-implementability of a social choice function as a truthful equilibrium outcome in the canonical mechanism. For certain degrees of the agent's weights for reciprocity, it becomes attractive for an agent to punish the other, trading-off material gains for psychological gains, which destroys the truth-telling NFE.

6 Psychologically robust mechanisms

Are there "psychologically robust" mechanisms, which yield truth-telling as the unique equilibrium outcome, whatever the agent's sensitivity for reciprocity? In this section I analyze the implementation problem demanding that the equilibrium outcome is robust to any psychologically motivations by the agents, i.e., a NE should be robust to varying degrees of the individual weights $(\xi_{ij})_{i,j \in I, i \neq j}$. Example 3 of the previous section demonstrates that in a specific environment all equilibria of the canonical mechanism are psychologically robust equilibria and they have the desired outcome. However, in general, a social choice function is not psychologically robustly implementable in the canonical mechanism.

I establish the result that mechanisms, in which the rules prohibit a single agent to affect the equilibrium outcome, are psychologically robust, in the sense that any unanimous message profile is a psychologically robust equilibrium of the mechanism. I call such mechanisms to be "outcome-robust". However, outcome-robust mechanisms cannot handle the problem of multiple equilibria. As soon as robustness of the equilibrium outcome to unilateral deviations is required, a message profile that is truth-telling in one state continuous to be an equilibrium for all other state, and the equilibrium outcome need not be unique.

The next section presents a mechanism that was proposed by Eliaz (2002) for fault-tolerant implementation. This "fault-tolerant mechanism" is psychologically robust. Subsequently, Section 6.2 exhibits a general result for outcome-robust mechanisms.

6.1 Robustness and the fault-tolerant mechanism

Eliaz (2002) analyzes the implementation problem allowing that a subset of $k > 0$ agents may fail to play equilibrium strategies, and establishes that a social choice function is implementable in a fault-tolerant equilibrium. Fault-tolerant implementation requires robustness to k-deviations by the agents, which implies that those agents cannot affect material payoffs in equilibrium. This makes Eliaz's analysis interesting for the problem of psychologically robust implementation. A psychologically robust implementation requires the equilibrium to be robust to deviations of the agents due to reciprocity, which in turn implies that the agents cannot affect material payoffs in order to trade-off material gains for psychological ones.

Fault-tolerant mechanism (Eliaz, 2002) Let $\Phi^F = (M, g)$ denote the fault-tolerant mechanism: each agent $i \in I$ simultaneously proposes a message $m_i = (x^i, \theta^i, z^i)$, where $x \in X$, $\theta \in \Theta$ and z a non-negative integer. Let K denote a subset of k "faulty" agents, i.e., these agents behave randomly and may fail to achieve their optimal strategies, and neither the social planner nor the majority of non-faulty agents know the identity or exact number of the faulty agents. Suppose $0 < k < \frac{N}{2} - 1$. The outcome function g is defined by the following three rules:

(1) If at least $N-k$ agents send the same message $m_1 = ... = m_{N-k} = (x, \theta)$, and $x = f(\theta)$, then the outcome is $g(m) = x$.

(2) If exactly $N-k-1$ agents send the same message $m_1 = ... = m_{N-k-1} = (x, \theta)$, and $x = f(\theta)$, then the outcome is $g(m) = x$, except if *all* of the remaining $k+1$ agents announce $m_{N-k} = m_N = (x', \theta')$, with $m_j \neq m_i$, for $i \neq j$, and $i \in I \setminus K$, $j \in K$ with $K \subseteq I$ such that $|K| = k+1$, then

   (i) $g(m) = x'$ if for everyone of them $x \succeq_i^\theta x'$, $i \in K$;

   (ii) $g(m) = x$ otherwise.

(3) Otherwise an integer game is played and $g(m) = x_i$, where i is such that $z^i \geq z^j$ for all $i, j \in I$, $i \neq j$. In case of ties, $g(m) = x_i$, where x_i is the alternative chosen by agent i, who is the agent with the lowest index among those agents who announced the highest integer.

Rule (1) of Φ^F defines that one agent alone cannot challenge the majority view; rule (2) adds that only a minority of $k+1$ can change the outcome. These assumptions are in contrast to the canonical mechanism, that enables a single agent to challenge the

outcome.[14] If a single agent cannot change the outcome, then her intentions when choosing one message or another become irrelevant, which implies the equilibrium to be psychologically robust. However, any unanimous message profile in which the agents coordinate on an outcome associated with the social choice function is a PRE.

Proposition 5. *Consider the fault-tolerant mechanism Φ^F, for any $k \geq 1$, and $N \geq 5$.*[15] *Any unanimous message profile m with $m_i = (x, \theta)$, and $x = f(\theta)$ is a PRE.*

Proof. Fix an (arbitrary) state $\theta \in \Theta$. Suppose all agents $i \in I$ coordinate on the same message $m_i = (x, \theta)$, and fix a state $\theta \in \Theta$. Since $f(\theta) = x$, rule (1) applies and $g(m) = x$. Then the profile m is a PRE: suppose agent $i' \neq i$ deviates by announcing some alternative message $m_i^D \neq m_i$. Then still rule (1) applies and the outcome remains x. Rule (1) guarantees that the outcome is unaffected by any unilateral deviation, since it is enough that $(N-1)$ agents coordinate on the same message. Hence, m is an equilibrium, for every $\xi \in [0, \infty)^{N(N-1)}$. □

By the rules of Φ^F no agent can influence the outcome by a unilateral deviation. This implies that the agents' cannot respond to the behavior of others in a reciprocal way. Hence, any profile in which all N agents coordinate on the same outcome associated with the SCF f for that state θ is a PRE. If θ is the true state, than the equilibrium is a truthful psychologically robust equilibrium. However, if θ is not the true state, then the equilibrium is a untruthful psychologically robust equilibrium. Therefore, fault-tolerant mechanism has multiple psychologically robust equilibria, implying that it does not only yield the desired outcome.[16]

6.2 A general result

The fact that a mechanism does not allow a single agent to change the equilibrium outcome, implies that reciprocity concerns cannot influence equilibrium behavior, i.e., the mechanism rules out the possibility to punish unkind or to reward kind behavior. Therefore, any Nash equilibrium message profile, which induces the desired outcome in one state, continues to be an equilibrium in some other state, which,

[14] Note that for $k = 0$ the fault-tolerant mechanism is the same as the canonical mechanism.

[15] By definition, the k-faulty agents are the minority, $k < \frac{1}{2}N - 1$. By Maskin's Theorem 1 we need $N \geq 3$ agents. Hence, to allow for $k \geq 1$, we must have $N \geq 5$.

[16] Eliaz discusses the multiple equilibrium problem, stating that "if all the players are non-faulty and each plays according to an undesirable NE and each believes that all the other players are non-faulty as well, then no player would have an incentive to deviate..." (p. 596), so that the outcome of the mechanism would be a non-truthful equilibrium. However, the multiplicity problem is ruled out by the introduction of a new solution concept called *k-fault tolerant Nash equilibrium* (k-FTNE). k-FTNE requires robustness to k deviations of the agents, so that each agent best responds to the actions of the $(N-k)$ non-faulty agents, regardless of the identity and actions of the faulty agents. Hence, k-FTNE is a strong refinement of NE. With the concept of PRE, it has been shown that the problem of non-uniqueness remains.

in turn, implies that the equilibrium outcome need not be unique. The following proposition can be stated.[17]

Proposition 6. *Consider any mechanism Φ with a profile of NE $m^*(\theta)$ for all $\theta \in \Theta$. Let the social choice function f be induced by $m^*(\theta)$, i.e., $f(\theta) = g(m^*(\theta))$. If $\exists \theta'$ such that*

$$g(m^*(\theta')) = g(m_i', m_{-i}^*(\theta')), \quad \forall m_i' \in M_i, \forall i, \tag{1}$$

then $m^*(\theta')$ is a PRE in Γ^θ for all $\theta \in \Theta$.
Hence, if f is not trivial, i.e., $f(\theta') \neq f(\theta'')$ for some $\theta'' \in \Theta$, the equilibrium outcome of Φ is not unique.

Proof. Fix a mechanism Φ. Let $m^*(\theta')$ be a NE of the game $\Gamma^{\theta'}$ induced by Φ in state θ' with $f(\theta') = g(m^*(\theta')) = g(m_i', m_{-i}^*(\theta'))$, $\forall m_i'$, $\forall i$. Then $m^*(\theta')$ is clearly a PRE of $\Gamma^{\theta'}$. Consider now any other game Γ^θ induced by Φ in state θ, with $\theta \in \Theta$, $\theta \neq \theta'$. Suppose all agents coordinate on $m^*(\theta')$ in Γ^θ. Since it holds that $g(m^*(\theta')) = g(m_i', m_{-i}^*(\theta'))$, $\forall m_i' \in M_i$, $\forall i$, $m^*(\theta')$ is also a PRE of Γ^θ.
But since f is non-trivial, we have $g(m^*(\theta')) = f(\theta') \neq f(\theta'')$, for some $\theta'' \in \Theta$. Hence, the equilibrium outcome in $\Gamma^{\theta''}$ is not unique. □

Hence, condition (1) of Proposition 6 ensures that any NE of a mechanism Φ is a PRE of that mechanism Φ. However, uniqueness of the equilibrium outcome cannot be achieved. As soon as the possibility that agents can affect the equilibrium outcome is shut down, robustness is obtained, but also non-uniqueness.

7 Conclusion

In this study, I incorporated reciprocity preferences into the classic implementation problem with complete information. Reciprocity captures the agents' willingness to trade-off material payoffs and kindness sensations. I adopt Rabin's (1993) solution concept of fairness equilibrium, and introduce the notion of Nash fairness implementation. A social choice function is defined to be Nash fairness implementable, if in every state of the world, every Nash fairness equilibrium of the game induced by a mechanism, has the desirable outcome, where the desirable outcome is the one associated with the social choice function in the true state.
In specific settings, I apply the canonical mechanism, that has been used for Nash

[17] Bierbrauer and Netzer (2012) derive a related result by introducing an "insurance property" on the SCF, that constitutes – in a Bayesian setting – together with Bayes-Nash implementability, a sufficient condition on the SCF to be implementable in PRE. Note that they do not require uniqueness of the equilibrium outcome.

implementation (Repullo, 1987; Maskin, 1999), to the solution concept of Nash fairness equilibrium.

To sum up the results, I show in three specific settings that (i) a Nash implementable social choice function is not necessarily Nash fairness implementable in the canonical mechanism. A small preference for reciprocity can be enough to break down the truthful equilibrium outcome; (ii) there might be many unwanted equilibrium outcomes in the canonical mechanism when the agents' concern for reciprocity are large; and (iii) it is indeed possible to implement a social choice function in the canonical mechanism whatever the agents' concern for reciprocity. Hence, the canonical mechanism can be the appropriate tool for Nash fairness implementation of a social choice function, but it need not necessarily be the case if the agents' preferences for reciprocity are large.

I derive a property that together with Nash implementability is sufficient for a social choice function to be *not* Nash fairness implementable whatever the agents' concern for reciprocity. There will be always an upper-bound on the individual weight that agents' place on reciprocity, such that for weights above this upper-bound truth-telling will not be an equilibrium outcome, and there might be undesired equilibrium outcomes.

At last, I address the psychological robustness of general mechanisms. A mechanism has a psychologically robust equilibrium if a message profile constitutes a Nash equilibrium irrespective of how much kindness sensations affect the agents' behavior. I establish that any Nash equilibrium of a mechanism, in which the equilibrium outcome cannot be changed by a single agent, is also a psychologically robust equilibrium. However, these mechanisms cannot handle the problem of multiple equilbria, i.e., any message profile that is a truth-telling in one state continuous to be an equilibrium in all other states, such that the equilibrium outcomes need not be unique.

Acknowledgements

The author thanks Nick Netzer and Armin Schmutzler as well as the seminar participants in Zurich for helpful comments and suggestions. Financial support of the Swiss National Science Foundation is gratefully acknowledged.

Appendix

A.1 The canonical mechanism of Example 1

In the canonical mechanism $\Phi^C = (M, g)$ of Example 1, each agent $i \in \{1, 2, 3\}$ simultaneously sends a message $m_i = (x^i, \theta^i, z^i) \in M_i$, with $\Theta = \{\theta_1, \theta_2, \theta_3\}$ the set of states of the world, $X = \{x_1, x_2, x_3\}$ the set of alternatives, and $\mathbb{Z}^+ = \{0, 1, 2, ...\}$ the set of non-negative integers. With slight abuse of notation, I write $m_i = (x, \theta)$ whenever $m_i = m_j$ for some arbitrary integer z^i.
The outcome function g of Φ^C is defined by the following three rules:

(1) If all agents send the same message, i.e.,

  (i) if $m_1 = m_2 = m_3 = (x_1, \theta_1)$, then the outcome is $g(m) = x_1$;

  (ii) if $m_1 = m_2 = m_3 = (x_2, \theta_2)$, then the outcome is $g(m) = x_2$;

  (iii) if $m_1 = m_2 = m_3 = (x_3, \theta_3)$, then the outcome is $g(m) = x_3$.

(2) If two agents $i \neq i'$ sent the same message $m_i = m_{i'} = (x_j, \theta_j)$, and $x_j = f(\theta_j)$, $j \in \{1, 2, 3\}$, then the outcome is $g(m) = x_j$, except if the two agents announce

  (i) $m_i = m_{i'} = (x_1, \theta_1)$, and the dissident is

    * agent 1 who announces $m_1 = (x_3^1, \theta_2^1, z^1)$, then the outcome is $g(m) = x_3$;
    * agent 1 who announces $m_1 = (x_3^1, \theta_3^1, z^1)$, then the outcome is $g(m) = x_3$;
    * agent 1 who announces $m_1 = (x_2^1, \theta_3^1, z^1)$, then the outcome is $g(m) = x_2$;
    * agent 3 who announces $m_3 = (x_2^3, \theta_2^3, z^3)$, then the outcome is $g(m) = x_2$.

  (ii) $m_i = m_{i'} = (x_2, \theta_2)$ and the dissident is

    * agent 3 who announces $m_3 = (x_1^3, \theta_3^3, z^3)$, then the outcome is $g(m) = x_1$;
    * agent 3 who announces $m_3 = (x_1^3, \theta_1^3, z^3)$, then the outcome is $g(m) = x_1$;
    * agent 3 who announces $m_3 = (x_3^3, \theta_1^3, z^3)$, then the outcome is $g(m) = x_3$;
    * agent 2 who announces $m_2 = (x_3^2, \theta_3^2, z^2)$, then the outcome is $g(m) = x_3$.

  (iii) $m_i = m_{i'} = (x_3, \theta_3)$ and the dissident is

    * agent 2 who announces $m_2 = (x_2^2, \theta_1^2, z^2)$, then the outcome is $g(m) = x_2$;
    * agent 2 who announces $m_2 = (x_2^2, \theta_2^2, z^2)$, then the outcome is $g(m) = x_2$;
    * agent 2 who announces $m_2 = (x_1^2, \theta_2^2, z^2)$, then the outcome is $g(m) = x_1$;
    * agent 1 who announces $m_1 = (x_1^1, \theta_1^1, z^1)$, then the outcome is $g(m) = x_1$.

(3) In all other cases, an integer game is played and the outcome is the alternative chosen by the agent who announces the highest integer. In case of ties, the outcome is the alternative chosen by the agent with the lowest index among those agents who announced the highest integer.

A.2 The canonical mechanism of Example 3

In the canonical mechanism $\Phi^C = (M, g)$ of Example 3, each agent $i \in \{1, 2, 3\}$ simultaneously sends a message $m_i = (x^i, \theta^i, z^i) \in M_i$, with $\Theta = \{\theta_1, \theta_2\}$ the set of states of the world, $X = \{x_1, x_2, x_3\}$ the set of alternatives, and $\mathbb{Z}^+ = \{0, 1, 2, ...\}$ the set of non-negative integers. With slight abuse of notation, I write $m_i = (x, \theta)$ whenever $m_i = m_j$ for some arbitrary integer z^i.
The outcome function g of Φ^C is defined by the following three rules:

(1) If all agents send the same message, i.e.,

   (i) if $m_1 = m_2 = m_3 = (x_1, \theta_1)$, then the outcome is $g(m) = x_1$;

   (ii) if $m_1 = m_2 = m_3 = (x_2, \theta_2)$, then the outcome is $g(m) = x_2$;

(2) If two agents $i \neq i'$ sent the same message $m_i = m_{i'} = (x_j, \theta_j)$, and $x_j = f(\theta_j)$, $j \in \{1, 2\}$, then the outcome is $g(m) = x_j$, except if the two agents announce

   (i) $m_i = m_{i'} = (x_1, \theta_1)$, and the dissident is

      * agent 1 who announces $m_1 = (x_2^1, \theta_2^1, z^1)$, then the outcome is $g(m) = x_2$;
      * agent 1 who announces $m_1 = (x_3^1, \theta_2^1, z^1)$, then the outcome is $g(m) = x_3$;

   (ii) $m_i = m_{i'} = (x_2, \theta_2)$ and the dissident is

      * agent 1 who announces $m_1 = (x_1^1, \theta_1^1, z^1)$, then the outcome is $g(m) = x_1$;
      * agent 1 who announces $m_1 = (x_3^1, \theta_1^1, z^1)$, then the outcome is $g(m) = x_3$;

(3) In all other cases, an integer game is played and the outcome is the alternative chosen by the agent who announces the highest integer. In case of ties, the outcome is the alternative chosen by the agent with the lowest index among those agents who announced the highest integer.

A.3 Proofs

In the following, with slight abuse of notation, I write $m_i = (x, \theta)$ whenever $m_i = m_j$ for some arbitrary integer z^i.

A.3.1 Monotonicity and NVP in Example 1

Claim 1 (Monotonicity). *The SCF f of Example 1 satisfies monotonicity.*

Proof. Fix two states $\theta_k, \theta_{k'} \in \Theta$, $k \neq k'$, and $k, k' \in \{1, 2, 3\}$. By definition $f(\theta_k) = x_k$. Since the outcome x_k is excluded from f when moving from state θ_k to another state $\theta_{k'}$, there must be preference reversal for at least one agent. Consider agent $i \in I$. We have $x_k \succ_i^{\theta_k} x_{k'}$, but $x_k \prec_i^{\theta_{k'}} x_{k'}$, which satisfies the requirement of preference reversal. □

Claim 2 (NVP). *The SCF f of Example 1 satisfies NVP.*

Proof. Since in each state $\theta_k \in \Theta$, $k \in \{1, 2, 3\}$, each agent top-ranks a different alternative, and $f(\theta_k) = x_k$, NVP is trivially satisfied. □

A.3.2 Monotonicity and NVP in Example 3

Claim 3 (Monotonicity). *The SCF f of Example 3 satisfies monotonicity.*

Proof. By definition $f(\theta_k) = x_k$, $k \in \{1, 2\}$, and outcome x_k is excluded from f when moving from state θ_1 to state θ_2 (or vice versa). Then there must be preference reversal for at least one agent. Consider agent 1. We have $x_1 \succ_1^{\theta_1} x_2$, but $x_1 \prec_1^{\theta_2} x_2$, which satisfies the requirement of preference reversal. q.e.d. □

Claim 4 (NVP). *The SCF f of Example 3 satisfies NVP.*

Proof. In each state $\theta_k \in \Theta$, $k \in \{1, 2\}$, $(N-1)$ agents top-rank the same alternative x_k, and $f(\theta_k) = x_k$, such that NVP is satisfied. q.e.d. □

A.3.3 Proof of Proposition 1

The proof of Proposition 1 will proceed through a number of lemmas. For convenience, I denote the message of agent i by $m_i = (x_j^i, \theta_k^i, z^i)$, where $i, j, k \in \{1, 2, 3\}$. For purpose of clarity, let θ_1 be the true state – the symmetry among agent's preferences can later be exploited to generalize the results for each of the three states being the true state.

Lemma 1. *Let θ_1 be the true state. Under the truth-telling message profile m^* with $m_i^* = (x_1^i, \theta_1^i, z^i) \in M_i$, for all pairs of agents, at least one of them is kindness neutral in equilibrium, hence*

$$U_1(m^*) = \pi_1(m^*) = H, \quad U_2(m^*) = \pi_2(m^*) = L, \quad U_3(m^*) = \pi_3(m^*) = M.$$

Proof. Suppose all three agents $i \in \{1, 2, 3\}$ announce the truth, $m_i^* = (x_1, \theta_1)$. Since $f(\theta_1) = x_1$, the outcome is x_1. The corresponding monetary payoffs of the agents are $\pi_1 = H$, $\pi_2 = L$, and $\pi_3 = \delta H + (1-\delta)L = M$. In order to determine the reciprocity payoff, first the equitable payoffs and kindness terms of all agents are computed, maintaining bilaterally Pareto efficiency within the set of feasible strategies. The equitable payoff, $\pi_{i'}^{e_i}((m_{ii'}^*)_{i' \neq i})$, for agent i' from the perspective of agent i, given truth-telling of the others with $i, i' \in \{1, 2, 3\}$ is

$$\begin{aligned}
\pi_1^{e_2}(m_1^*, m_3^*) &= \frac{1}{2}[H + H] = H; \\
\pi_1^{e_3}(m_1^*, m_2^*) &= \frac{1}{2}[H + H] = H; \\
\pi_2^{e_1}(m_2^*, m_3^*) &= \frac{1}{2}[H + L]; \\
\pi_2^{e_3}(m_2^*, m_1^*) &= \frac{1}{2}[H + L]; \\
\pi_3^{e_1}(m_3^*, m_2^*) &= \frac{1}{2}[H + M]; \\
\pi_3^{e_2}(m_3^*, m_1^*) &= \frac{1}{2}[M + M] = M.
\end{aligned}$$

E.g., consider $\pi_2^{e_1}(m_2^*, m_3^*)$ and $\pi_3^{e_1}(m_3^*, m_2^*)$. Agent 1 (since she is a test agent) potentially can induce either outcome x_3 or x_2 (see rule 2 (i) in Appendix A.1). Alternative x_2 (x_3) is top-ranked for agent 2 (3), which would result in her highest material payoff, respectively. Since both x_2 and x_3 make agent 1 herself worse off, the lowest payoff agent 1 can give to agent 3 is M; agent 2 already receives her lowest payoff $= L$. Since the equitable payoff is the average between the minimum and maximum payoff agent 1 can potentially give to each agent, we have $\pi_2^{e_1}(m_2^*, m_3^*) = \frac{1}{2}[H + L]$ and $\pi_3^{e_1}(m_3^*, m_2^*) = \frac{1}{2}[H + M]$.

Agent i's belief about her own kindness towards agent i', $\kappa_{ii'}(m_i, m_{-i}) = \pi_{i'}(m_i, m_{-i}) - \pi_{i'}^{e_i}((m^*_{ii'})_{i' \neq i})$ can be calculated as follows:

$$\begin{aligned}
&\kappa_{12}(m^*) = \lambda_{212}(m^*) = L - \frac{1}{2}[H + L] = \frac{1}{2}(L - H) < 0;\\
&\kappa_{13}(m^*) = \lambda_{313}(m^*) = M - \frac{1}{2}[H + M] = \frac{1}{2}(1 - \delta)(L - H) < 0;\\
&\kappa_{21}(m^*) = \lambda_{121}(m^*) = H - H = 0;\\
&\kappa_{23}(m^*) = \lambda_{323}(m^*) = M - M = 0;\\
&\kappa_{31}(m^*) = \lambda_{131}(m^*) = H - H = 0;\\
&\kappa_{32}(m^*) = \lambda_{232}(m^*) = L - \frac{1}{2}(H + L) = \frac{1}{2}(L - H) < 0.
\end{aligned}$$

Then, the fairness utility for each agent i, $U_i(m^*) = \pi_i(m^*) + \sum_{i \neq i'} \xi_{ii'} \kappa_{ii'}(m^*) \lambda_{ii'i}(m^*)$ is given as

$$U_1(m^*) = \pi_1(m^*) = H, \quad U_2(m^*) = \pi_2(m^*) = L, \quad U_3(m^*) = \pi_3(m^*) = M.$$

The fairness utilities coincide with the (pure) material payoffs, i.e., the fairness components drop out of the agents' utilities, because in each pair at least one agent is kindness neutral. □

Lemma 2. *Let θ_1 be the true state. Given the truth-telling message profile m^*, there exists a message m_i^D for agent 1 and agent 3, but not for agent 2, such that $U_i(m_i^D, m^*_{-i}) \neq U_i(m^*)$. In particular, for agent 1*

- *$U_1(m_1^D, m_2^*, m_3^*) = L$, if $m_1^D = (x_3^1, \theta_2^1, z^1)$;*
- *$U_1(m_1^D, m_2^*, m_3^*) = L$, if $m_1^D = (x_3^1, \theta_3^1, z^1)$;*
- *$U_1(m_1^D, m_2^*, m_3^*) = M$, if $m_1^D = (x_2^1, \theta_2^1, z^1)$;*

and for agent 3

- *$U_3(m_1^*, m_2^*, m_3^D) = L + \xi_{31} \frac{1}{4}(M - H)^2$, if $m_3^D = (x_2^3, \theta_2^3, z^3)$.*

Proof. Given the truth-telling message $m_i^* = (x_1^i, \theta_1^i, z^i)$ of agent $i \in \{1, 2, 3\}$. Consider a unilateral deviation of agent i from m_i^*.

(i) *Deviation by agent 1.* Consistent with rule 2 of Φ^C (see Appendix A.1), agent 1 has three options to change the outcome of Φ^C:

(1) If agent 1 sends the message $m_1^{D1} = (x_3^1, \theta_2^1, z^1)$, then the outcome of Φ^C changes to $g(m_1^{D1}, m^*_{-i}) = x_3$. Outcome x_3 results in a strict worsening compared to the truth-telling outcome for agent 1, $\pi_1(m_1^{D1}, m^*_{-i}) = L < H$, but a strict improvement for both, agent 2 and 3, who receive $\pi_2(m_1^{D1}, m^*_{-i}) = M$

and $\pi_3(m_1^{D1}, m_{-i}^*) = H$, respectively. Determining kindness, we have

$$\kappa_{12}(m_1^{D1}, m_{-i}^*) = M - \frac{1}{2}(H+L) = (\delta - \frac{1}{2})(H-L) > 0, \quad \text{if} \quad \delta > \frac{1}{2};$$
$$\kappa_{13}(m_1^{D1}, m_{-i}^*) = H - \frac{1}{2}(H+M) = \frac{1}{2}(1-\delta)(H-L) > 0.$$

Note that agent 1's belief about her kindness towards agent 2 does depend on δ. That is, not only the smaller the spread between the best and intermediate outcome, or the larger the spread between the intermediate and worst outcome (the higher δ), the more agent 1 beliefs to be kind towards agent 2. For the critical value $\delta = \frac{1}{2}$, agent 1 is kindness neutral towards agent 2; as soon as $\delta > \frac{1}{2}$ agent 1 thinks to be kind to agent 2, otherwise she thinks to be unkind.

(2) If agent 1 sends the message $m_1^{D2} = (x_3^1, \theta_3^1, z^1)$, then the outcome of Φ^C changes to $g(m_1^{D2}, m_{-i}^*) = x_3$. Payoffs and kindness terms are similar to those in (1).

(3) If agent 1 sends the message $m_1^{D3} = (x_2^1, \theta_2^1, z^1)$, then the outcome changes to $g(m_1^{D3}, m_{-i}^*) = x_2$. Outcome x_2 results in a strict worsening for both, agent 1 and agent 3, $\pi_1(m_1^{D3}, m_{-i}^*) = M < H$ and $\pi_3(m_1^{D3}, m_{-i}^*) = L < M$, respectively, and in a strict improvement for agent 2, $\pi_2(m_1^{D3}, m_{-i}^*) = H$. The corresponding kindness terms are

$$\kappa_{12}(m_1^{D3}, m_{-i}^*) = H - \frac{1}{2}(H+L) = \frac{1}{2}(H-L) > 0;$$
$$\kappa_{13}(m_1^{D3}, m_{-i}^*) = L - \frac{1}{2}(H+L) = \frac{1}{2}(L-H) < 0.$$

In all cases (1)-(3), agent 1's beliefs regarding the kindness of agent 2 and 3 do not change and are similar to the ones under truth-telling, i.e., $\lambda_{121}(m^*) = \lambda_{131}(m^*) = 0$. Hence, agent 1's fairness utility after deviation from truth-telling is

$$U_1^D(m_1^D, m_{-i}^*) = \begin{cases} L, & \text{if (1) or (2);} \\ M, & \text{if (3).} \end{cases}$$

Agent 1's fairness utility after deviation is independent of any reciprocity components and coincides with the material payoff. Since $U_1(m^*) = H > U_1(m_1^D, m_2^*, m_3^*)$, there does not exist a "profitable" deviation from truth-telling by agent 1.

(ii) *Deviation by agent 2.* Agent 2 receives her worst alternative from truth-telling. By rule 2 of Φ^C (see Appendix A.1) she cannot change the outcome by a unilateral deviation. Thus, $U_2(m_1^*, m_2^D, m_3^*) = U_2(m^*) = L$.

(iii) *Deviation by agent 3.* Suppose agent 3 unilaterally deviates from truth-telling by announcing $m_3^D = (x_2^3, \theta_2^3, z^3)$. Then, by rule 2 of Φ^C (see Appendix A.1), the outcome changes to x_2. Outcome x_2 results in monetary payoffs of $\pi_1(m_3^D, m_{-i}^*) = M$, $\pi_2(m_3^D, m_{-i}^*) = H$, and $\pi_3(m_3^D, m_{-i}^*) = L$. This is a strict worsening for agent 1

and 3, but a strict improvement for agent 2. The kindness terms are

$$\kappa_{31}(m_3^D, m_{-i}^*) = M - \frac{1}{2}[H + M] = \frac{1}{2}(1-\delta)(L-H) < 0;$$
$$\kappa_{32}(m_3^D, m_{-i}^*) = H - \frac{1}{2}[H + L] = \frac{1}{2}(H-L) > 0.$$

Further, $\lambda_{131}(m^*) = \frac{1}{2}(1-\delta)(L-H) < 0$, $\lambda_{232}(m^*) = 0$, $\lambda_{313}(m^*) = \frac{1}{2}(1-\delta)(L-H) < 0$, and $\lambda_{323}(m^*) = 0$. Hence, we have

$$U_3(m_3^D, m_{-i}^*) = L + \xi_{31}\frac{1}{4}(1-\delta)^2(H-L)^2 > L \quad \text{if} \quad \xi_{31} > 0.$$

Here, for some $\xi_{31} > 0$, agent 3 receives a positive reciprocity payoff in addition to her material payoff.

□

Lemma 3. *Let θ_1 be the true state. Then,*

(i) agent 1 and 2 never have an incentive to deviate from truth-telling, for any $\xi \in [0,\infty)^6$;

(ii) agent 3 has no incentive to deviate from truth-telling if and only if $0 \leq \xi_{31} \leq \varepsilon_1$ with $\varepsilon_1 = \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)}$.

Hence, the truth-telling message profile m^* is a NFE if and only if $\xi_{31} \leq \varepsilon_1$.

Proof. Comparing fairness utilities, agents 1 and 2 both have no incentive to deviate from truth-telling, since $U_1(m_1^D, m_{-i}^*) < U_1(m^*)$ and $U_2(m_2^D, m_{-i}^*) = U_2(m^*)$, respectively. However, agent 3 has an incentive to deviate from truth-telling for certain values of ξ_{31}. That is,

$$\begin{aligned} U_3(m_3^D, m_{-i}^*) \geq U_3(m^*) \quad &\iff \quad L + \xi_{31}\frac{1}{4}(1-\delta)^2(L-H)^2 \geq M \\ &\iff \quad \xi_{31} \geq \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)}. \end{aligned}$$

Agent 3 does not deviate from truth-telling if and only if $0 \leq \xi_{31} \leq \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)}$. Thus, there is an upper-bound on agent 3's reciprocity weight ξ_{31}. If $\xi_{31} > \frac{\delta}{(1-\delta)^2}\frac{4}{(H-L)}$, then m^* is not a NFE. □

Lemma 4. *Let θ_1 be the true state and consider the message profile m with $m_1 = m_2 = m_3 = (x_k, \theta_k)$, and $k \in \{2,3\}$. Then m is not a NFE, for all $\xi \in [0,\infty)^6$.*

Proof. Suppose all three agents untruthfully announce the same message, $m \neq m^*$ with $m_i = (x_k, \theta_k)$, $k \in \{2,3\}$. Then rule 1 of Φ^C (see Appendix A.1) applies and $g(m) = f(\theta_k) = x_j$. For agent 1 or agent 3 the outcome results in the lowest material payoff.

Focus on agent 3. If $m_i = (x_2, \theta_2)$ for all agents $i \in I$, then $\pi_3(m) = L$, $\pi_2(m) = H$, $\pi_1(m) = M$. The relevant kindness terms for agent 3 are

$$\begin{aligned}
\kappa_{31}(m) &= \lambda_{131}(m) = M - \frac{1}{2}(H+L) = (\delta - \frac{1}{2})(H-L);\\
\kappa_{32}(m) &= \lambda_{232}(m) = H - \frac{1}{2}(H+L) = \frac{1}{2}(H-L);\\
\kappa_{13}(m) &= \lambda_{313}(m) = L - L = 0;\\
\kappa_{23}(m) &= \lambda_{323}(m) = L - \frac{1}{2}(H+L) = \frac{1}{2}(L-H).
\end{aligned}$$

Then agent 3's fairness utility can be calculated as

$$U_3(m) = L - \xi_{32}\frac{1}{4}(H-L)^2.$$

Unilateral deviation A potential deviation strategy of agent 3 is to announce $m_3^D = (x_3^3, \theta_3^3, z^3)$. Given the fixed strategies of agent 1 and 2, $m_1 = m_2 = (x_2, \theta_2)$, rule 2 of Φ^C applies and the outcome of Φ^C changes to x_3. Material payoffs are $\pi_1(m_3^D, m_{-i}) = L$, $\pi_2(m_3^D, m_{-i}) = M$, and $\pi_3(m_3^D, m_{-i}) = H$. Whereas agent 3's beliefs regarding the kindness of agent 1 and 2 do not change, $\lambda_{313}(m) = 0$, and $\lambda_{323}(m) = \frac{1}{2}(L-H) < 0$, agent 3's own kindness towards the others does,

$$\begin{aligned}
\kappa_{31}(m_3^D, m_{-i}) &= L - \frac{1}{2}(H+L) = \frac{1}{2}(L-H);\\
\kappa_{32}(m_3^D, m_{-i}) &= M - \frac{1}{2}(H+L) = (H-L)(\delta - \frac{1}{2}).
\end{aligned}$$

Then agent 3's fairness utility after deviation is

$$U_3(m_3^D, m_{-i}) = H - \xi_{32}\frac{1}{2}(\delta - \frac{1}{2})(H-L)^2.$$

Comparing fairness utilities, it follows

$$U_3(m_3^D, m_{-i}) > U_3(m) \quad \Longleftrightarrow \quad \frac{\delta - 1}{2} < \frac{1}{\xi_{32}(H-L)}.$$

The result will apply similarly if the focus is on agent 1, assuming $m_1 = m_2 = m_3 = (x_3, \theta_3)$. Hence, the profile $m = ((x_2^1, \theta_2^1, z^1), (x_2^2, \theta_2^2, z^2), (x_2^3, \theta_2^3, z^3))$ as well as the profile $m = ((x_3^1, \theta_3^1, z^1), (x_3^2, \theta_3^2, z^2), (x_3^3, \theta_3^3, z^3))$ cannot be NFE. □

Lemma 5. *Let θ_1 be the true state and consider a message profile m with $m_i = m_{i'} \neq m_{i''}$, such that rule 2 of Φ^C applies. Then m is not a NFE, for all $\xi \in [0, \infty)^6$.*

Proof. Suppose two agents $i \neq i'$ announce the same message, $m_i = m_{i'} = (x_k, \theta_k)$, with $k \in \{1, 2, 3\}$, such that $f(\theta_k) = x_k$, but agent i'' announces $m_{i''} \neq m_i$, such that the outcome of Φ^C will be determined by rule 2 (see Appendix A.1). For each possible strategy profile it needs to be proven that at least one agent has an incentive to unilaterally deviate

from her announced message.
In the following all possible strategy profiles and corresponding outcomes are listed. Note that in brackets on the right side the rank of the resulting outcome for a corresponding agent are identified. In particular, the left entry identifies the outcome to result in the intermediate material payoff, M, or in the worst material payoff, L, for a particular agent i, identified by the right entry. This classification will be used later in the proof.

(1) $m_1 = m_2 \neq m_3$ with

(a)	$m_1 = m_2 = (x_1, \theta_1)$	and	$m_3 = (x_2^3, \theta_3^3, z^3)$	$\rightarrow$	$g(m) = x_2$	[M,1]
(b)	$m_1 = m_2 = (x_1, \theta_1)$	and	$m_3 \neq (x_2^3, \theta_3^3, z^3)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(c)	$m_1 = m_2 = (x_2, \theta_2)$	and	$m_3^{(1)} = (x_1^3, \theta_3^3, z^3)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(d)	$m_1 = m_2 = (x_2, \theta_2)$	and	$m_3^{(2)} = (x_1^3, \theta_1^3, z^3)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(e)	$m_1 = m_2 = (x_2, \theta_2)$	and	$m_3^{(3)} = (x_3^3, \theta_1^3, z^3)$	$\rightarrow$	$g(m) = x_3$	[L,1]
(f)	$m_1 = m_2 = (x_2, \theta_2)$	and	$m_3 \neq m_3^{(1)}, m_3^{(2)}, m_3^{(3)}$	$\rightarrow$	$g(m) = x_2$	[L,3]
(g)	$m_1 = m_2 = (x_3, \theta_3)$	and	m_3 arbitrary	$\rightarrow$	$g(m) = x_3$	[L,1]

(2) $m_1 = m_3 \neq m_2$ with

(a)	$m_1 = m_3 = (x_2, \theta_2)$	and	$m_2 = (x_3^2, \theta_3^2, z^2)$	$\rightarrow$	$g(m) = x_2$	[L,3]
(b)	$m_1 = m_3 = (x_2, \theta_2)$	and	$m_2 \neq (x_3^2, \theta_3^2, z^2)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(c)	$m_1 = m_3 = (x_3, \theta_3)$	and	$m_2^{(1)} = (x_2^2, \theta_1^2, z^2)$	$\rightarrow$	$g(m) = x_2$	[L,3]
(d)	$m_1 = m_3 = (x_3, \theta_3)$	and	$m_2^{(2)} = (x_1^2, \theta_2^2, z^2)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(e)	$m_1 = m_3 = (x_3, \theta_3)$	and	$m_2^{(3)} = (x_2^2, \theta_2^2, z^2)$	$\rightarrow$	$g(m) = x_2$	[L,3]
(f)	$m_1 = m_3 = (x_3, \theta_3)$	and	$m_2 \neq m_2^{(1)}, m_2^{(2)}, m_2^{(3)}$	$\rightarrow$	$g(m) = x_3$	[L,1]
(g)	$m_1 = m_3 = (x_1, \theta_1)$	and	m_2 arbitrary	$\rightarrow$	$g(m) = x_1$	[M,3]

(3) $m_2 = m_3 \neq m_1$ with

(a)	$m_2 = m_3 = (x_3, \theta_3)$	and	$m_1 = (x_1^1, \theta_1^1, z^1)$	$\rightarrow$	$g(m) = x_1$	[L,2]
(b)	$m_2 = m_3 = (x_3, \theta_3)$	and	$m_1 \neq (x_1^1, \theta_1^1, z^1)$	$\rightarrow$	$g(m) = x_3$	[L,1]
(c)	$m_2 = m_3 = (x_1, \theta_1)$	and	$m_1^{(1)} = (x_3^1, \theta_2^1, z^1)$	$\rightarrow$	$g(m) = x_3$	[L,1]
(d)	$m_2 = m_3 = (x_1, \theta_1)$	and	$m_1^{(2)} = (x_2^1, \theta_3^1, z^1)$	$\rightarrow$	$g(m) = x_2$	[L,3]
(e)	$m_2 = m_3 = (x_1, \theta_1)$	and	$m_1^{(3)} = (x_3^1, \theta_3^1, z^1)$	$\rightarrow$	$g(m) = x_3$	[L,1]
(f)	$m_2 = m_3 = (x_1, \theta_1)$	and	$m_1 \neq m_1^{(1)}, m_1^{(2)}, m_1^{(3)}$	$\rightarrow$	$g(m) = x_1$	[L,2]
(g)	$m_2 = m_3 = (x_2, \theta_2)$	and	m_1 arbitrary	$\rightarrow$	$g(m) = x_2$	[L,3]

Given the strategy profiles, it can be shown that in each case at least one agent has an incentive to deviate from her announcement. Since the incentive to deviate depends on the outcome the agents receive, two cases can be separated, focussing on the agents who (i) get the intermediate material payoff (cases (1a) and (2g)), or (ii) get the lowest material payoff (all remaining cases). The particular agent has an incentive to unilaterally deviate, eliciting the integer game. This is shown by an example for each of the two cases.

Case (i): Profile (1a). Suppose the strategy profile $m = ((x_1^1, \theta_1^1, z^1), (x_1^2, \theta_1^2, z^2), (x_2^3, \theta_2^3, z^3))$ is a NFE. The mechanism's outcome is $g(m) = x_2$, which results in the material payoffs $\pi_1(m) = M$, $\pi_2(m) = H$, and $\pi_3(m) = L$, i.e., agent 1 gets the intermediate payoff. Calculating the equitable payoffs, one need to be careful to guarantee bilateral efficiency among strategies. It is always possible for the agent to induce the integer game such that the maximum (potential) payoff, that she could have given to the other agents, coincides with their best payoff. But the minimum (potential) payoff must not always equal the worst payoff, if this would make both agents worse off in a bilateral Pareto sense. Straightforward calculation then lead to the relevant equitable payoff for agent 1, that are

$$\pi_2^{e_1}(m_2, m_3) = \frac{1}{2}[H + L]; \quad \pi_3^{e_1}(m_3, m_2) = \frac{1}{2}[H + L];$$
$$\pi_1^{e_2}(m_1, m_3) = \frac{1}{2}[H + M]; \quad \pi_1^{e_3}(m_1, m_2) = \frac{1}{2}[H + M].$$

Then, agent 1's kindness towards agents 2 and 3 is

$$\kappa_{12}(m) = \frac{1}{2}(H - L) > 0, \quad \text{and} \quad \kappa_{13}(m) = \frac{1}{2}(L - H) < 0,$$

with corresponding beliefs about the kindness of agent 2 and 3

$$\lambda_{121}(m) = \frac{1}{2}(1 - \delta)(L - H) < 0, \quad \text{and} \quad \lambda_{131}(m) = \frac{1}{2}(1 - \delta)(L - H) < 0.$$

Hence, agent 1's fairness utility is

$$U_1(m) = M - (H - L)^2 \frac{1}{4}(1 - \delta)[\xi_{12} + \xi_{13}].$$

Suppose now agent 1 unilaterally deviates by announcing $m_1^D = (x_1^1, \theta_1^1, z^1)$ with $z^1 > z^2, z^3$. Then the outcome - by the integer game - changes to x_1, resulting in $\pi_1^D = H$. Kindness is determined as

$$\kappa_{12}(m_1^D, m_{-i}) = \frac{1}{2}(L - H), \quad \text{and} \quad \kappa_{13}(m_1^D, m_{-i}) = (\delta - \frac{1}{2})(H - L).$$

Then agent 1's fairness utility after deviation is

$$U_1^D(m_1^D, m_{-i}) = H + (H - L)^2 \frac{1}{2}(1 - \delta)[\frac{1}{2}\xi_{12} - (\delta - \frac{1}{2})\xi_{13}],$$

which always exceeds her fairness utility before,

$$U_1(m_1^D, m_{-i}) > U_1(m) \quad \Longleftrightarrow \quad H - M > -(H - L)^2 \frac{1}{2}(1 - \delta)[\xi_{12} + \xi_{13}(1 - \delta)].$$

Case (ii): Profile (1d). Suppose the strategy profile $m = ((x_2^1, \theta_2^1, z^1), (x_2^2, \theta_2^2, z^2), (x_1^3, \theta_1^3, z^3))$ is a NFE. The outcome of the mechanism is $g(m) = x_1$, and material payoffs are $\pi_1(m) = H$, $\pi_2(m) = L$, and $\pi_3(m) = M$, i.e., agent 2 receives the lowest payoff. The relevant

equitable payoffs for agent 2 are

$$\pi_1^{e_2}(m_1, m_3) = \pi_3^{e_2}(m_1, m_3) = \pi_2^{e_1}(m_2, m_3) = \pi_2^{e_3}(m_1, m_2) = \frac{1}{2}(H + L).$$

Then, agent 2's kindness towards agents 1 and 3 is

$$\kappa_{21}(m) = \frac{1}{2}(H - L), \quad \text{and} \quad \kappa_{23}(m) = (\delta - \frac{1}{2})(H - L) > 0,$$

and her associated beliefs about the others kindness towards herself are

$$\lambda_{212}(m) = \frac{1}{2}(H - L) \quad \text{and} \quad \lambda_{232}(m) = \frac{1}{2}(L - H).$$

Hence, agent 2's fairness utility is calculated as

$$U_2(m) = L - (H - L)^2[\frac{1}{4}\xi_{21} + \frac{1}{2}(\delta - \frac{1}{2})\xi_{23}].$$

However, agent 2 can do better. Given $m_1 = (x_2^1, \theta_2^1, z^1)$, and $m_3 = (x_1^3, \theta_1^3, z^3)$, agent 2 has an incentive to deviate to $m_2^D = (x_2^2, \theta_2^2, z^2)$ with $z^2 > z^1, z^3$. Then the integer game takes place and agent 2 wins in the sense that the outcome changes to be x_2 (and $\pi_2(m_2^D, m_{-i}) = H$). Hence,

$$\kappa_{21}(m_2^D, m_{-i}) = \frac{1}{2}(1 - \delta)(L - H), \quad \text{and} \quad \kappa_{23}(m_2^D, m_{-i}) = \frac{1}{2}(H - L).$$

Then agent 2's fairness utility after deviation is

$$U_2(m_2^D, m_{-i}) = H + (H - L)^2[\frac{1}{4}(1 - \delta)\xi_{21} + \frac{1}{4}\xi_{23}].$$

which always exceeds her fairness utility before, since

$$U_2(m_2^D, m_{-i}) > U_2(m) \quad \Longleftrightarrow \quad 1 > -(H - L)[\frac{1}{4}\xi_{21}(2 - \delta) + \frac{1}{2}\delta\xi_{23}].$$

□

Lemma 6. *Let θ_1 be the true state. Any message profile m such that rule (3) applies is not a NFE, for all $\xi \in [0, \infty)^6$.*

Proof. Suppose all three agents announce different states and/or outcomes, such that the integer game takes play. By the cyclical order of preferences among agents, there will be always one agent i for whom the outcome is bad in the sense that it results in her lowest material payoff. In her eyes at least one of the other agent's was unkind, since they could have announced a lower integer or agent i's favorite outcome. Suppose $m_1 \neq m_2 \neq m_3$ and $m_1 = (x_1^1, \theta_1^1, z^1)$ with $z^1 > z^2, z^3$. By rule 3 of Φ^C (see Appendix A.1), agent 1 wins the integer game, and outcome x_1 is implemented, which is the one that yields the highest material payoff for herself, $\pi_1(m) = H$. Consider agent 2 for whom outcome x_1 yields the

lowest material payoff, $\pi_2(m) = L$, and

$$\kappa_{12}(m) = \frac{1}{2}(L - H) < 0, \quad \text{and} \quad \kappa_{32}(m) = \frac{1}{2}(L - H) < 0.$$

However, agent 2 is kind towards agent 1, and unkind towards agent 3 iff $\delta < \frac{1}{2}$,

$$\begin{aligned}\kappa_{21}(m) &= \frac{1}{2}(H - L) > 0,\\ \kappa_{23}(m) &= (\delta - \frac{1}{2})(H - L) < 0, \quad \text{iff} \quad \delta < \frac{1}{2}.\end{aligned}$$

Hence, agent 2's fairness utility is determined as

$$U_2(m) = L - \xi_{21}\frac{1}{4}(H - L)^2 - \xi_{23}\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2.$$

Consider the unilateral deviation $m_2^D = (x_2^2, \theta_1^2, z^2)$, with $z^2 > z^1, z^3$, by agent 2 such that her favorite outcome x_2 is implemented. Then,

$$\begin{aligned}\kappa_{23}(m_2^D, m_{-i}) &= \frac{1}{2}(L - H) < 0,\\ \kappa_{21}(m_2^D, m_{-i}) &= (\delta - \frac{1}{2})(H - L) < 0, \quad \text{iff} \quad \delta < \frac{1}{2}.\end{aligned}$$

The fairness utility after deviation is

$$U_2^D(m_2^D, m_{-i}) = H - \xi_{21}\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2 + \xi_{23}\frac{1}{4}(H - L)^2,$$

which strictly exceeds the one before, $U_2(m_2^D, m_{-i}) > U_2(m)$. Hence, agent 2 always has an incentive to deviate in the integer game. Note that the result does not depend on whether the agent who wins the integer game actually announced an alternative that yields the lowest material payoff for herself. E.g., suppose that $m_1 \neq m_2 \neq m_3$ and $m_1 = (x_3^1, \theta_1^1, z^1)$ with $z^1 > z^2, z^3$, such that outcome x_3 will be implemented by rule 3 of Φ^C. Again, agent 1 wins the integer game, but the outcome x_3 actually yields the lowest material payoff for herself, $\pi_1(m) = L$. By the same logic as above, agent 1 always has an incentive to deviate, announcing the message $m_1 = (x_1^1, \theta_1^1, z^1)$ with $z^1 > z^2, z^3$ such that her favorite outcome will be implemented.
The results hold for each agent and state. Thus, there is no equilibrium of the integer game. □

A.3.4 Proof of Proposition 2

Suppose θ_1 is the true state. If all agents coordinate on the same outcome associated with the SCF by announcing $m_i^* = (x_1, \theta_1)$, $i \in I$, then the outcome is x_1. Since x_1 is top-ranked by all agents, and since strategies must be bilaterally Pareto efficient, we have $\kappa_{ii'}(m_i^*, m_{-i}^*) = 0$ for all $i, i' \in I$, $i \neq i'$, which implies $U_i(m^*) = \pi_i(m^*) = H$. No agent has an incentive to deviate to any other message $m_i^D \neq m_i^*$, since by deviation $\pi_i^D(m_i^D, m_{-i}^D) < \pi_i(m^*)$ and since $\lambda_{ii'i}(m^*) = 0$ for all $i, i' \in I, i \neq i'$, so that any deviation will result in a fairness utility lower than the fairness utility from truth-telling: $U_i^D(m_i^D, m_{-i}^*) = \pi_i^D(m_i^D, m_{-i}^D) < U_i(m^*)$. Thus, the truth-telling message profile m^* is a NFE, for every $\xi \in [0, \infty)^6$.

Now, assume that all agents coordinate on the same but untruthful message profile $\tilde{m} \neq m^*$, with $\tilde{m}_i = (x_2, \theta_2)$, such that each agent announces her least preferred alternative (in the true state θ_1). Since $f(\theta_2) = x_2$, rule (1) applies and the outcome is x_2, which yields the minimum material payoff for every agent, $\pi_i(\tilde{m}) = L$. All agents are as unkind as possible towards one another, since the equitable payoff coincides with the maximum material payoff, i.e., the minimum payoff agent i can give to another agent i' is H, otherwise strategies are not bilaterally Pareto efficient. In particular, we have $\pi_{i'}^{e_i}(m_i, m_{i''}) = \frac{1}{2}(H + H) = H$, and $\kappa_{ii'}(\tilde{m}) = (L - H) < 0$, for all $i, i' \in I$, $i \neq i'$. Hence, the fairness utility is $U_i(\tilde{m}) = L + \sum_{i' \neq i} \xi_{ii'}(H - L)^2$.

Consider two possible deviation strategies. First, consider the possible unilateral deviation $\tilde{m}_i^{D1} = (x_3^i, \theta_2^i, z^i)$ by agent $i \in I$, such that the outcome will change to x_3. By this deviation, agent i is unkind towards both other agents, $\kappa_{ii'}(\tilde{m}_i^{D1}, \tilde{m}_{-i}) = \kappa_{ii''}(\tilde{m}_i^{D1}, \tilde{m}_{-i}) = (1 - \delta)(L - H) < 0$. Given beliefs, agent i's fairness utility becomes $U_i(\tilde{m}_i^{D1}, \tilde{m}_{-i}) = M + \sum_{i' \neq i} \xi_{ii'}(1 - \delta)(H - L)^2$, for all $i, i' \in I$, $i \neq i'$. Comparing fairness utilities, we have

$$U_i(\tilde{m}) \geq U_i(\tilde{m}_i^{D1}, \tilde{m}_{-i}) \quad \text{iff} \quad \sum_{i' \neq i} \xi_{ii'} \geq \frac{1}{(H - L)}.$$

Second, consider the possible unilateral deviation $\tilde{m}_i^{D2} = (x_1^i, \theta_1^i, z^i)$ by agent i, such that the outcome will change to x_1. By this deviation, agent i is now kindness-neutral towards agent i' and agent i'', i.e., by the same reason as before, the equitable payoff coincides with the highest material payoff, hence $\kappa_{ii'}(\tilde{m}_i^{D2}, \tilde{m}_{-i}) = \kappa_{ii''}(\tilde{m}_i^{D2}, \tilde{m}_{-i}) = H - \frac{1}{2}(H + H) = 0$. Given beliefs, agent i's fairness utility now becomes $U_i(\tilde{m}_i^{D2}, \tilde{m}_{-i}) = H$, for all $i, i' \in I$, $i \neq i'$. Again, comparing fairness utilities we have

$$U_i(\tilde{m}) \geq U_i(\tilde{m}_i^{D2}, \tilde{m}_{-i}) \quad \text{iff} \quad \sum_{i' \neq i} \xi_{ii'} \geq \frac{1}{(H - L)}.$$

Hence, iff $\sum_{i' \neq i} \xi_{ii'} \geq \frac{1}{(H - L)}$, then $\tilde{m}$ is a NFE.

□

A.3.5 Proof of Proposition 3

The proof of Proposition 3 will proceed through a number of lemmas. Again, for purpose of clarity let θ_1 represent the true state. The symmetry among agent's preferences allows us to generalize the results for θ_2 being the true state. Note that the intermediate payoff M is defined for $\delta \in (\frac{1}{2}, 1)$.

Lemma 7. *Let θ_1 be the true state. Then the truth-telling message profile m^* with $m_i^* = (x_1^i, \theta_1^i, z^i) \in M_i$ is a PRE.*

Proof. W.o.l.g. let x^H denote the alternative which is top-ranked in an agent's preference ordering, and x^L the alternative that is lowest-ranked. If all agents announce the truth, $m_i^* = (x_1, \theta_1)$, for all $i \in \{1, 2, 3\}$, then rule 1 of Φ^C (see Appendix A.2) applies and the outcome is $g(m^*) = x_1$, with $x_1 = x^H$ for agent 1 and 2, and $x_1 = x^L$ for agent 3; further, we have $\kappa_{12}(m^*) = \kappa_{21}(m^*) = \kappa_{31}(m^*) = \kappa_{32}(m^*) = 0$, implying $U_i(m^*) = \pi_i^*(m^*)$, for all agents $i \in \{1, 2, 3\}$, in particular $U_1(m^*) = U_2(m^*) = H$, and $U_3(m^*) = L$. Then, no agent has an incentive to deviate from truth-telling for all $\xi_{ii'} \in [0, \infty)^6$. Suppose agent i announces $m_i^D \neq m^*$. Since beliefs are $\lambda_{121}(m^*) = \lambda_{131}(m^*) = \lambda_{212}(m^*) = \lambda_{232}(m^*) = 0$, we have $U_1(m_1^D, m_2^*, m_3^*) = \pi_1(m_1^D, m_2^*, m_3^*) < U_1(m^*)$, and $U_2(m_1^*, m_2^D, m_3^*) = \pi_2(m_1^*, m_2^D, m_3^*) < U_2(m^*)$. Consider therefore agent 3. However, since $x_1 = x^L$ for agent 3, she cannot change the equilibrium outcome by the rules of Φ^C. Hence, for any message $m_3^D \neq m_3^*$, we have $\kappa_{31}(m_1^*, m_2^*, m_3^D) = \kappa_{32}(m_1^*, m_2^*, m_3^D) = 0$, and thus $U_3(m_3^D, m_1^*, m_2^*) = \pi_3(m_3^D, m_1^*, m_2^*) = L = U_3(m^*)$. □

Lemma 8. *Let θ_1 be the true state and consider the message profile m with $m_1 = m_2 = m_3 = (x_2, \theta_2)$. Then m is not a NFE, for all $\xi \in [0, \infty)^6$.*

Proof. Suppose all three agents untruthfully announce the same message, $m \neq m^*$ with $m_i = (x_2, \theta_2)$, such that rule 1 of Φ^C (see Appendix A.2) applies. Then $g(m) = f(\theta_2) = x_2$, and material payoffs are $\pi_1(m) = L$, $\pi_2(m) = M$, and $\pi_3(m) = H$. The kindness terms are determined as $\kappa_{12}(m) = (\delta - \frac{1}{2})(H - L) > 0$, $\kappa_{13}(m) = \frac{1}{2}(H - L) > 0$, and $\kappa_{21}(m) = \kappa_{23}(m) = \kappa_{31}(m) = \kappa_{32}(m) = 0$. This implies that each agent's fairness utility equals the material payoff, i.e., $U_i(m) = \pi_i(m)$.
Consider now a possible deviation by agent 1 to $m_1^D = (x_1^1, \theta_1^1, z^1)$, such that the outcome will change to x_1. Since alternative x_1 results in a material payoff of $\pi_1(m_1^D, m_2, m_3) = H$ for agent 1, and since beliefs are zero, it holds that $U_1(m_1^D, m_2, m_3) = \pi_1(m_1^D, m_2, m_3) = H > L = \pi_1(m) = U_1(m)$, implying that the profile m cannot be a NFE. □

Lemma 9. *Let θ_1 be the true state and consider a message profile m with $m_i = m_{i'} \neq m_{i''}$ such that rule 2 of Φ^C applies. Then, any profile $\tilde{m}$ with $\tilde{m}_1 = \tilde{m}_2 = (x_1, \theta_1) \neq \tilde{m}_3$ is a PRE with outcome $g(\tilde{m}) = g(m^*) = x_1$.*

Proof. Suppose two agents $i \neq i'$ announce the same message, $m_i = m_{i'} = (x_k, \theta_k)$, $k \in \{1, 2\}$, but the third agent i'' announces $m_{i''} \neq m_i$ such that rule 2 of Φ^C (see Appendix A.2 applies. In particular, the possible strategy profiles and corresponding outcomes are

(a) $m_i = m_{i'} = (x_1, \theta_1) \neq m_{i''}$

(1)	$m_1 = m_2 \neq m_3$	and	m_3 arbitrary	$\rightarrow$	$g(m) = x_1$
(2)	$m_1 = m_3 \neq m_2$	and	m_2 arbitrary	$\rightarrow$	$g(m) = x_1$
(3)	$m_2 = m_3 \neq m_1$	and	$m_1^{(1)} = (x_2^1, \theta_2^1, z^1)$	$\rightarrow$	$g(m) = x_2$
(4)	$m_2 = m_3 \neq m_1$	and	$m_1^{(2)} = (x_3^1, \theta_2^1, z^1)$	$\rightarrow$	$g(m) = x_3$
(5)	$m_2 = m_3 \neq m_1$	and	$m_1 \neq m_1^{(1)}, m_1^{(2)}$	$\rightarrow$	$g(m) = x_1$

(b) $m_i = m_{i'} = (x_2, \theta_2) \neq m_{i''}$

(1)	$m_1 = m_2 \neq m_3$	and	m_3 arbitrary	$\rightarrow$	$g(m) = x_2$
(2)	$m_1 = m_3 \neq m_2$	and	m_2 arbitrary	$\rightarrow$	$g(m) = x_2$
(3)	$m_2 = m_3 \neq m_1$	and	$m_1^{(1)} = (x_1^1, \theta_1^1, z^1)$	$\rightarrow$	$g(m) = x_1$
(4)	$m_2 = m_3 \neq m_1$	and	$m_1^{(2)} = (x_3^1, \theta_1^1, z^1)$	$\rightarrow$	$g(m) = x_3$
(5)	$m_2 = m_3 \neq m_1$	and	$m_1 \neq m_1^{(1)}, m_1^{(2)}$	$\rightarrow$	$g(m) = x_2$

Case (a). Consider the case when two agents announce the same message $m_i = m_{i'} = (x_1, \theta_1) \neq m_{i''}$. In particular,

(1) if $m_1 = m_2 = (x_1, \theta_1) \neq m_3$, with m_3 arbitrary, then the outcome is x_1. The situation is similar to the one in Lemma 7. Agent 3 cannot change the outcome by the rules of Φ^C, such that it does not make a difference which strategy she chooses. Hence, fairness utilities from m also coincide with the material payoffs for each agent; and no agent has an incentive to unilaterally deviate from m for each $\xi_{ii'} \in [0, \infty)^6$, since $U_1(m_1^D, m_2, m_3) = \pi_1(m_1^D, m_2, m_3) < U_1(m)$, and $U_2(m_1, m_2^D, m_3) = \pi_2(m_1, m_2^D, m_3) < U_2(m)$, and $U_3(m_3^D, m_1, m_2) = \pi_3(m_3, m_1, m_2) = L = U_3(m)$. Thus, the outcome under m coincides with the one under the truth-telling equilibrium message profile m^*.

(2) if $m_1 = m_3 = (x_1, \theta_1) \neq m_2$, with m_2 arbitrary, then the outcome is x_1. Material payoffs are $\pi_1(m) = \pi_2(m) = H$, and $\pi_3(m) = L$; and kindness terms are $\kappa_{12}(m) = 0$, $\kappa_{13}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{21}(m) = 0$, $\kappa_{23}(m) = \frac{1}{2}\delta(L - H) < 0$, $\kappa_{31}(m) = \kappa_{32}(m) = \frac{1}{2}(H - L) > 0$. Consider agent 3 with

$$U_3(m) = L - \xi_{31}\frac{1}{4}(H - L)^2 - \xi_{32}\frac{1}{4}\delta(H - L)^2.$$

Then agent 3 has an incentive to deviate to $m_3^D = (x_2^3, \theta_2^3, z^3)$, with $z^3 > z^1, z^2$, inducing the integer game such that the outcome changes to x_2: $\pi_3(m_3^D, m_{-i}) = H$, and $\kappa_{31}(m_3^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, and $\kappa_{32}(m_3^D, m_{-i}) = (\delta - \frac{1}{2})(H - L) > 0$, such that

$$U_3(m_1, m_2, m_3^D) = H + \xi_{31}\frac{1}{4}(H - L)^2 - \xi_{32}\frac{1}{2}\delta(\delta - \frac{1}{2})(H - L)^2 > U_3(m).$$

(3) if $m_2 = m_3 = (x_1, \theta_1)$, but $m_1 = (x_2^1, \theta_2^1, z^1)$, then the outcome is x_2. Outcome x_2 results in material payoffs $\pi_1(m) = L$, $\pi_2(m) = M$, and $\pi_3(m) = H$; and kindness terms are $\kappa_{12}(m) = (\delta - \frac{1}{2})(H - L) > 0$, $\kappa_{13}(m) = \frac{1}{2}(L - H) < 0$,

$\kappa_{21}(m) = \frac{1}{2}(L-H) < 0$, $\kappa_{23}(m) = \frac{1}{2}(H-L) > 0$, $\kappa_{31}(m) = \frac{1}{2}(L-H) < 0$, $\kappa_{32}(m) = \frac{1}{2}(1-\delta)(H+L) > 0$. Consider agent 1 with

$$U_1(m) = L - \xi_{12}\frac{1}{2}(\delta - \frac{1}{2})(H-L)^2 + \xi_{13}\frac{1}{4}(H-L)^2.$$

Then agent 1 has an incentive to deviate to $m_1^D = (x_1^1, \theta_1^1, z^1)$, such that the outcome changes to x_1: $\pi_1(m_1^D, m_{-i}) = H$, and $\kappa_{12}(m_1^D, m_{-i}) = 0$, and $\kappa_{13}(m_1^D, m_{-i}) = \frac{1}{2}(L-H) < 0$, such that

$$U_1(m_1^D, m_2, m_3) = H + \xi_{13}\frac{1}{4}(H-L)^2 > U_1(m).$$

(4) if $m_2 = m_3 = (x_1, \theta_1)$, but $m_1 = (x_3^1, \theta_2^1, z^1)$, then the outcome is x_3. Outcome x_3 results in material payoffs $\pi_1(m) = \pi_3(m) = M$, and $\pi_2(m) = L$; and kindness terms are $\kappa_{12}(m) = \frac{1}{2}(L-H) < 0$, $\kappa_{13}(m) = \kappa_{21}(m) = \kappa_{23}(m) = \kappa_{31}(m) = (\delta - \frac{1}{2})(H-L) > 0$, and $\kappa_{32}(m) = \frac{1}{2}(L-H) < 0$. Consider agent 2 with

$$U_2(m) = L - (\xi_{21} + \xi_{23})\frac{1}{2}(\delta - \frac{1}{2})(H-L)^2.$$

Then agent 2 has an incentive to deviate to $m_2^D = (x_1^2, \theta_1^2, z^2)$, with $z^2 > z^1, z^2$ such that the integer game takes place and the outcome changes to x_1: $\pi_2(m_2^D, m_{-i}) = H$, and $\kappa_{21}(m_2^D, m_{-i}) = 0$, and $\kappa_{23}(m_2^D, m_{-i}) = \frac{1}{2}(L-H) < 0$, such that

$$U_2(m_1, m_2^D, m_3) = H + \xi_{32}\frac{1}{4}(H-L)^2 > U_2(m).$$

(5) if $m_2 = m_3 = (x_1, \theta_1) \neq m_1$, with $m_1 \neq (x_2^1, \theta_2^1, z^1)$, and, $m_1 \neq (x_3^1, \theta_2^1, z^1)$, then the outcome is x_1. Material payoffs are $\pi_1(m) = \pi_2(m) = H$, and $\pi_3(m) = L$; and kindness terms are $\kappa_{12}(m) = 0$, $\kappa_{13}(m) = \frac{1}{2}(L-H) < 0$, $\kappa_{21}(m) = 0$, $\kappa_{23}(m) = \frac{1}{2}(L-H) < 0$, $\kappa_{31}(m) = \kappa_{32}(m) = \frac{1}{2}(H-L) > 0$. Consider agent 3 with

$$U_3(m) = L - (\xi_{31} + \xi_{23})\frac{1}{4}(H-L)^2.$$

Then agent 3 has an incentive to deviate to $m_3^D = (x_2^3, \theta_2^3, z^3)$, with $z^3 > z^1, z^2$, inducing the integer game such that the outcome changes to x_2: $\pi_3(m_3^D, m_{-i}) = H$, and $\kappa_{31}(m_3^D, m_{-i}) = \frac{1}{2}(L-H) < 0$, and $\kappa_{32}(m_3^D, m_{-i}) = (\delta - \frac{1}{2})(H-L) > 0$, such that

$$U_3(m_1, m_2, m_3^D) = H + \xi_{31}\frac{1}{4}(H-L)^2 - \xi_{32}\frac{1}{2}(\delta - \frac{1}{2})(H-L)^2 > U_3(m).$$

Case (b). Consider the case when two agents announce the same message $m_i = m_{i'} = (x_2, \theta_2) \neq m_{i''}$. In particular,

(1) if $m_1 = m_2 = (x_2, \theta_2) \neq m_3$, with m_3 arbitrary, then the outcome is x_2. Material payoffs are $\pi_1(m) = L$, $\pi_2(m) = M$, and $\pi_1(m) = H$; and kindness terms are

$\kappa_{12}(m) = (\delta - \frac{1}{2})(H - L) > 0$, $\kappa_{13}(m) = \frac{1}{2}(H - L) > 0$, $\kappa_{21}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{23}(m) = \frac{1}{2}(H - L) > 0$, and $\kappa_{31}(m) = \kappa_{32}(m) = 0$. Consider agent 1 with

$$U_1(m) = L - \xi_{12}\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2.$$

Then agent 1 has an incentive to deviate to $m_1^D = (x_1^1, \theta_1^1, z^1)$ such that the outcome changes to x_1: $\pi_1(m_1^D, m_{-i}) = H$, and $\kappa_{12}(m_1^D, m_{-i}) = 0$, and $\kappa_{13}(m_1^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, such that

$$U_1(m_1^D, m_2, m_3) = H > U_1(m).$$

(2) if $m_1 = m_3 = (x_2, \theta_2) \neq m_2$, with m_2 arbitrary, then the outcome is x_2. Material payoffs are $\pi_1(m) = L$, $\pi_2(m) = M$, and $\pi_1(m) = H$; and kindness terms are $\kappa_{12}(m) = (\delta - \frac{1}{2})(H - L) > 0$, $\kappa_{13}(m) = \frac{1}{2}(H - L) > 0$, $\kappa_{21}(m) = \kappa_{23}(m) = 0$, $\kappa_{31}(m) = \frac{1}{2}(L - H) < 0$, and $\kappa_{32}(m) = \frac{1}{2}(1 - \delta)(L - H) < 0$. Consider agent 1 with

$$U_1(m) = L - \xi_{13}\frac{1}{4}(H - L)^2.$$

Then agent 1 has an incentive to deviate to $m_1^D = (x_1^1, \theta_1^1, z^1)$ such that the outcome changes to x_1: $\pi_1(m_1^D, m_{-i}) = H$, and $\kappa_{12}(m_1^D, m_{-i}) = 0$, and $\kappa_{13}(m_1^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, such that

$$U_1(m_1^D, m_2, m_3) = H + \xi_{13}\frac{1}{4}(H - L)^2 > U_1(m).$$

(3) if $m_2 = m_3 = (x_2, \theta_2)$, but $m_1 = (x_1^1, \theta_1^1, z^1)$, then the outcome is x_1. Outcome x_1 results in material payoffs $\pi_1(m) = \pi_2(m) = H$, and $\pi_3(m) = L$; and kindness terms are $\kappa_{12}(m) = 0$, $\kappa_{13}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{21}(m) = 0$, $\kappa_{23}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{31}(m) = \frac{1}{2}(H - L) > 0$, $\kappa_{32}(m) = \frac{1}{2}(H - L) > 0$. Consider agent 3 with

$$U_3(m) = L - (\xi_{32} + \xi_{32})\frac{1}{4}(H - L)^2.$$

Then agent 3 has an incentive to deviate to $m_3^D = (x_2^3, \theta_1^3, z^3)$, with $z^3 > z^1, z^2$ such that the outcome changes to x_2 by the integer game (rule (3)): $\pi_3(m_3^D, m_{-i}) = H$, and $\kappa_{31}(m_3^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, and $\kappa_{32}(m_3^D, m_{-i}) = (\delta - \frac{1}{2})(H - L) > 0$, such that

$$U_3(m_1, m_2, m_3^D) = H + \xi_{31}\frac{1}{4}(H - L)^2 - \xi_{32}\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2 > U_3(m).$$

(4) if $m_2 = m_3 = (x_2, \theta_2)$, but $m_1 = (x_3^1, \theta_1^1, z^1)$, then the outcome is x_3. Outcome x_1 results in material payoffs $\pi_1(m) = \pi_3(m) = M$, and $\pi_2(m) = L$; and kindness terms are $\kappa_{12}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{13}(m) = \kappa_{21}(m) = \kappa_{23}(m) = \kappa_{31}(m) =$

$(\delta - \frac{1}{2})(H - L) > 0$, and $\kappa_{32}(m) = \frac{1}{2}(L - H) < 0$. Consider agent 2 with

$$U_2(m) = L - (\xi_{21} + \xi_{23})\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2.$$

Then agent 2 has an incentive to deviate to $m_2^D = (x_1^2, \theta_1^2, z^2)$, with $z^2 > z^1, z^3$ such that the outcome changes to x_1 by the integer game (rule (3)): $\pi_2(m_2^D, m_{-i}) = H$, and $\kappa_{21}(m_2^D, m_{-i}) = 0$, and $\kappa_{23}(m_2^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, such that

$$U_2(m_1, m_2^D, m_3) = H + \xi_{23}\frac{1}{4}(H - L)^2 > U_2(m).$$

(5) if $m_2 = m_3 = (x_2, \theta_2) \neq m_1$, with $m_1 \neq (x_1^1, \theta_1^1, z^1)$, and $m_1 \neq (x_3^1, \theta_1^1, z^1)$, then the outcome is x_2. Material payoffs are $\pi_1(m) = L$, $\pi_2(m) = M$, and $\pi_1(m) = H$; and kindness terms are $\kappa_{12}(m) = (\delta - \frac{1}{2})(H - L) > 0$, $\kappa_{13}(m) = \frac{1}{2}(H - L) > 0$, $\kappa_{21}(m) = \frac{1}{2}(L - H) < 0$, $\kappa_{23}(m) = \frac{1}{2}(H - L) > 0$, $\kappa_{31}(m) = \frac{1}{2}(L - H) < 0$, and $\kappa_{32}(m) = \frac{1}{2}(1 - \delta)(L - H) < 0$. Consider agent 1 with

$$U_1(m) = L - \xi_{12}\frac{1}{2}(\delta - \frac{1}{2})(H - L)^2 - \xi_{13}\frac{1}{4}(H - L)^2.$$

Then agent 1 has an incentive to deviate to $m_1^D = (x_1^1, \theta_1^1, z^1)$ such that the outcome changes to x_1: $\pi_1(m_1^D, m_{-i}) = H$, and $\kappa_{12}(m_1^D, m_{-i}) = 0$, and $\kappa_{13}(m_1^D, m_{-i}) = \frac{1}{2}(L - H) < 0$, such that

$$U_1(m_1^D, m_2, m_3) = H + \xi_{13}\frac{1}{4}(H - L)^2 > U_1(m).$$

□

Lemma 10. *Let θ_1 be the true state. Any profile such that rule (3) applies is not a NFE.*

Proof. Suppose all three agents announce different messages $m_i \neq m_{i'} \neq m_{i''}$ such that the integer game takes place. Then the outcome by this message profile m results always in the lowest material payoff for one of the agents; and this agent will always have an incentive to deviate. Consider step by step all three possible outcomes:
first, suppose the outcome of the integer game is x_1. Then $\pi_1(m) = \pi_2(m) = H$ and $\pi_3(m) = L$; the kindness terms are $\kappa_{12}(m) = \kappa_{21} = 0$, $\kappa_{13}(m) = \kappa_{23}(m) = \frac{1}{2}(L - H) < 0$, and $\kappa_{31}(m) = \kappa_{32}(m) = \frac{1}{2}(H - L) > 0$. Agent 3 is the agent for whom outcome x_1 results in her lowest material payoff; hence, the fairness utility is

$$U_3(m) = L - (\xi_{31} + \xi_{32})\frac{1}{4}(H - L)^2.$$

Consider a unilateral deviation by agent 2 such that her favorite outcome x_2 is implemented, that is $m_3^D = (x_2, \theta_1, z^3)$ with $z^3 > z^1, z^2$. Then $\kappa_{31}(m_3^D, m_{-i}) = \kappa_{32}(m_3^D, m_{-i}) =$

$\frac{1}{2}(L-H)<0$, and

$$U_3(m_3^D, m_2, m_3) = H + (\xi_{31}+\xi_{32})\frac{1}{4}(H-L)^2 > U_3(m).$$

Second, suppose the outcome of the integer game is x_2. Then $\pi_1(m)=L$, $\pi_2(m)=M$, and $\pi_3(m)=H$; the kindness terms are $\kappa_{12}(m)=(\delta-\frac{1}{2})(H-L)$, $\kappa_{13}(m)=\frac{1}{2}(H-L)>0$, $\kappa_{21}(m)=\frac{1}{2}(L-H)<0$, $\kappa_{23}(m)=\frac{1}{2}(H-L)>0$, $\kappa_{31}(m)=\frac{1}{2}(L-H)<0$, and $\kappa_{32}(m)=\frac{1}{2}(1-\delta)(L-H)<0$. Agent 1 is now the agent for whom outcome x_2 results in her lowest material payoff; and her fairness utility is

$$U_1(m) = L - \xi_{12}\frac{1}{2}(\delta-\frac{1}{2})(H-L)^2 - \xi_{13}\frac{1}{4}(H-L)^2.$$

Consider a unilateral deviation by agent 1 to $m_1^D=(x_1^1,\theta_1^1,z^1)$ with $z^1>z^2,z^3$ such that the outcome changes to x_1. Then $\kappa_{12}(m_1^D,m_{-i})=0$ and $\kappa_{13}(m_1^D,m_{-i})=\frac{1}{2}(L-H)<0$, and

$$U_1^D(m_1^D, m_2, m_3) = H + \xi_{13}\frac{1}{4}(H-L)^2 > U_1(m).$$

Finally, if the outcome is x_3, material payoffs are $\pi_1(m)=M$, $\pi_2(m)=L$, and $\pi_3(m)=M$, and it is now agent 2 for whom x_2 results in the lowest material payoff. The kindness terms are $\kappa_{12}(m)=\frac{1}{2}(L-H)>0$, $\kappa_{13}(m)=\kappa_{21}(m)=\kappa_{23}(m)=\kappa_{31}(m)=(\delta-\frac{1}{2})(H-L)$, and $\kappa_{32}(m)=\frac{1}{2}(L-H)<0$; and agent 2's fairness utility is

$$U_2(m) = L - (\xi_{21}+\xi_{23})\frac{1}{2}(\delta-\frac{1}{2})(H-L)^2.$$

Consider a unilateral deviation by 2 to $m_2^D=(x_1^2,\theta_1^2,z^2)$, with $z^2>z^1,z^3$, such that the outcome changes to x_1. Then $\kappa_{21}(m_2^D,m_{-i})=0$, and $\kappa_{23}(m_2^D,m_{-i})=\frac{1}{2}(L-H)<0$, and

$$U_2(m_1, m_2^D, m_3) = H + \xi_{23}\frac{1}{4}(H-L)^2 > U_2(m).$$

□

References

Battigalli, P., and M. Dufwenberg (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1–35.

Ben-Porath, E., and B. L. Lipman (2012): "Implementation with partial provability," *Journal of Economic Theory*, in Press.

Bierbrauer, F., and N. Netzer (2012): "Mechanism Design and Intentions," *Working Paper*.

Bolton, G. E., and A. Ockenfels (2000): "ERC - A Theory of Equity, Reciprocity and Competition," *American Economic Review*, 90(1), 166–193.

Cabrales, A. (1999): "Adaptive Dynamics and the Implementation Problem with Complete Information," *Journal of Economic Theory*, 86(2), 159–184.

Cabrales, A., G. Charness, and L. C. Corchon (2003): "An Experiment on Nash Implementation," *Journal of Economic Behavior and Organization*, 51, 161–193.

Cabrales, A., and R. Serano (2011): "Implementation in adaptive better-response dynamics: Towards a general theory of bounded rationality in mechanisms," *Games and Economic Behavior*, 73, 360–374.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *The Quarterly Journal of Economics*, 117(3), 817–869.

Cox, J., D. Friedman, and S. Gjerstad (2007): "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*, 59, 17–45.

Danilov, V. (1992): "Implementation via Nash Equilibrium," *Economcetrica*, 60, 43–56.

Doghmi, A., and A. Ziad (2009): "Faulty Nash Implementation in Exchange Economies with Single-Peaked Preferences," *Working Paper*.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268–298.

Dutta, B., and A. Sen (1991): "A Necessary and Sufficient Condition for Two-Person Nash Implementation," *Review of Economic Studies*, 58(121-128).

Dutta, B., and A. Sen (2011): "Nash Implementation with Partially Honest Individuals," *Working paper*.

Eliaz, K. (2002): "Fault Tolerant Implementation," *Review of Economic Studies*, 69, 589–610.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114(3), 817–868.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Reciprocity," *Games and Economic Behavior*, 1, 60–79.

Jackson, M. O. (1992): "Implementation in Undominated Strategies: A Look at Bounded Mechanisms," *The Review of Economic Studies*, 59(4), 757–775.

——— (2001): "A Crash Course in Implementation Theory," *Social Choice and Welfare*, 18, 655–708.

Kartik, N., and O. Tercieux (2011): "Implementation with Evidence," *Working Paper*.

Levine, D. (1998): "Modelling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1, 593–622.

Lombardi, M., and N. Yoshilhara (2011): "Partially-Honest Nash Implementation: Characterization Results," *Working Paper*.

Mas-Colell, A., M. D. Winston, and J. R. Green (1995): *Microeconomic Theory*. Oxford University Press: USA.

Maskin, E. (1999): "Nash Equilibrium and Welfare Optimality," *Review of Economic Studies*, 66, 23–38.

Matushima, H. (2008a): "Behavioral Aspects of Implementation Theory," *Economic Letters*, 100, 161–164.

——— (2008b): "Role of Honesty in Full Implementation," *Journal of Economic Theory*, 139, 353–359.

Moore, J. (1992): "Implementation, Contracts and Renegotiation in Environments with Complete Information," in *Advances in Economic Theory: Invited papers for the Sixth World Congress of the Econometric Society*, ed. by J.-J. Laffont, vol. 1, pp. 182–282. Cambridge University Press.

Moore, J., and R. Repullo (1988): "Subgame Perfect Implementation," *Econometrica*, 58, 1083–1099.

Moore, J., and R. Repullo (1990): "Nash Implementation: A Full Characterization," *Econometrica*, 58, 1083–1099.

Rabin, M. (1993): "Incorporating Fairness into Games Theory and Economics," *American Economic Review*, 83, 1281–1302.

Repullo, R. (1987): "A Simple Proof of Maskin's Theorem on Nash-implementation," *Social Choice and Welfare*, 4, 39–41.

Saijo, T. (1988): "Strategy Space Reduction in Maskin's Theorem," *Econometrica*, 56, 693–700.

Segal, U., and J. Sobel (2007): "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory*, 136, 197–216.

Sjöström, T. (1991): "On the Necessary and Sufficient Conditions for Nash Implememtation," *Social Choice and Welfare*, 8, 333–340.

——— (1993): "Implementation in Perfect Equilibria," *Social Choice and Welfare*, 10, 97–106.

Sobel, J. (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 43, 392–436.

Wiliams, S. (1986): "Realization and Nash-implementation: Two Aspects of Mechanism Design," *Econometrica*, 54, 65–73.

Chapter 2

Cooperation, Communication, and Partner Selection in a High-Stakes Field Experiment

joint with Donja Darai

1 Introduction

A large body of evidence from experimental and empirical studies, performed by psychologists and economists, indicates that people are motivated by something other than only their material well-being. Communication is an important and powerful tool to influence individual behavior. Over the last decades, the impact of communication has been analyzed in many respects (see e.g., Crawford, 1998; Farrell and Rabin, 1996; Ledyard, 1995). One of the most prominent ideas is that cost-free communication comprises of more than non-binding cheap talk, especially in games where a selfish action conflicts with the socially optimal one. For instance, Sally (1995) and Ellingsen and Johannesson (2004) show that mere promises are effective in fostering cooperation and trust; and e.g., Charness and Dufwenberg (2006) and Gneezy (2005) propose that people show a reluctance to lie as a matter of guilt aversion or morals. However, non-verbal means of communication, like gestures and handshakes, as well as consequences of revealed lies on future behavior have so far - to the best of our knowledge - received no attention in the economic analysis.

In this study we investigate the impact of verbal and non-verbal communication, precommitment, and previous interactions on the formation of cooperative behavior.

We use field data from the British television game show "Golden Balls", in which at the end a slightly modified version of the prisoner's dilemma game is played.
The show starts with two rounds of pre-play, in which stakes are accumulated and in which two of initially four contestants are selected to proceed to the final round, the prisoner's dilemma. Stakes are accumulated via a random process, which does not involve the contestant's cognitive ability or an effort task. However, the stakes are partly private information to a contestant and are not observable to the others. This lack of information between contestants allows them to lie and bluff, but these lies are disclosed by the show host after each round. Therefore, after each pre-play round not only stakes are common knowledge among contestants, but also who was honest and who was not.
The selection among contestants takes place via a voting process, in which at the end of each pre-play round every contestant has to vote for one among them to leave the show. In the final stage the two remaining contestants play to eventually share the accumulated stakes in a prisoner's dilemma, which slightly differs from the standard model since defection is a weakly dominant strategy.
The game show distinguishes itself from laboratory experiments through extraordinary high stakes – the average stake size amounts to £13 000 – and face-to-face communication between contestants.
Our analysis shows a unilateral cooperation rate of 54% and contestants even manage to mutually cooperate in 33% of cases. Communication, both verbal and non-verbal, and stake size have a major impact on cooperative behavior. In particular, contestants who promise to cooperate and who in addition corroborate their intention to cooperate with a handshake are indeed significantly more likely to cooperate than if they make no promises and handshakes at all. However, a mere handshake is actually used to manipulate the opponent's attitude towards cooperation: contestants who shake hands immediately before they choose their action for the prisoner's dilemma, are actually 19 percentage points more likely to defect. Concerning stake size, we identify a negative correlation between stake size as well as expected stake size and cooperation. Also the propensity to lie, i.e., to neglect a promise or handshake, is significantly positively correlated with stake size.
The data also offers the opportunity to test for consistent behavior of contestants (e.g., Falk and Zimmermann, 2011). Immediately before the show, contestants are privately asked to state whether they intend to defect or cooperate in case they reach the final round. This statement is broadcasted to the television audience, but not to the other contestants. We test whether a contestant's pre-action statement can serve as a predictor for actual behavior. The results show that 75% (65%) of contestants who explicitly state to cooperate (defect) before the show, indeed cooperate (defect) when they actually choose their action in the prisoner's dilemma.
The particular structure of the pre-play allows us to investigate lying in a situation where people are aware that lies are revealed and individually identified. We find

that lying is punished and that contestants are bounded rational in their decision to lie. A liar of the current or previous round is significantly more often eliminated from the show than an honest contestant. For this reason lying is costly, but still more than half of the players lie in at least one of the pre-play rounds, and they lie not only when it is particularly harmful for them to be honest. And, contestants condition their lie and the size of overstatement merely on their own position in the game, but neglect the strengths and weaknesses of their opponents. Further, a liar from the pre-play encounters a reputation cost, which has an effect on the outcome of the prisoner's dilemma: two finalist who have been honest throughout the game, manage to mutually cooperate significantly more likely than contestants who have lied before; though, lying is no predictor for stealing behavior.
Besides, we control for several individual player characteristics, such as age, gender, race, place of residence, occupation, as well as for observational learning through experience in later episodes.

Our study is related to independent work by van den Assem et al. (2012), who analyze cooperative behavior using data from "Golden Balls".[1] However, the analysis of (non-)verbal communication and pre-commitment for cooperative behavior, as well as lying, and the partner selection process is exclusive to our study. To strengthen the explanatory power of our results, we control for their findings and can confirm their results with respect to player characteristics and stake size (e.g., "small peanut phenomenon"). However, we offer a different channel to identify reciprocal behavior and find support for conditional cooperation.

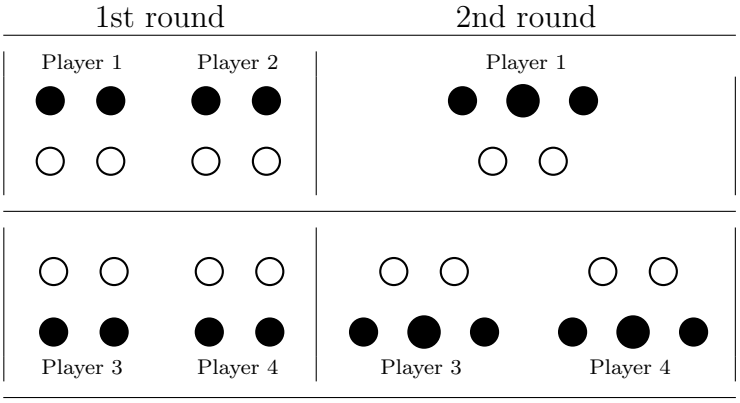
The remainder of this study is organized as follows. In the next section we describe the course of the game show and the data set. Section 3 refers to the analysis of communication and cooperation. Section 4 proceeds by the analysis of lying and partner selection in two stages of pre-play. Finally, Section 5 concludes.

2 Game show and data set

In this section we describe in detail the course of events in the game show (Section 2.1) and the data set (Section 2.2).

[1]While writing the first draft of this study, which was published in July 2010, it came to our attention that van den Assem et al. (2012) are independently analyzing data from "Golden Balls" (first working paper version in April 2010). Both data sets are independently and individually established, and they differ by the collection, definition and construction of variables.

Figure I: Structure of round 1 and 2



Note: Black balls depict "closed back row balls"; white balls depict "open front row balls".

2.1 Structure of the game show

The game show "Golden Balls" consists of three rounds of play with the final round being divided into two phases.

Round 1 The game show starts with four contestants, usually two women and two men, who are briefly introduced by the show host, i.e., the contestants provide some information about themselves including their names, occupation and place of residence.[2] Then the first round starts: 16 golden balls are mixed, twelve of them have written a cash amount (in £) inside and four have written the word "killer" inside. Killer balls are dangerous, because these may damage the jackpot in the final round. The balls containing a cash value are drawn from a lottery of 100 golden balls with a minimum ball value of £10 and a maximum ball value of £75 000.[3] Each contestant arranges the closed golden balls in two rows of two balls in front of herself (see Figure I, left). The two balls on the front row are opened by each contestant, thus their content is common knowledge. The content of the remaining two balls on the back row is private information to each contestant, i.e., the contestants are allowed to secretly look inside. Afterwards the show host asks each contestant to state what is inside her closed balls. The order in which contestants are asked for their claims is exogenously determined by the show host. A discussion between the contestants about the claims that have been made follows. The discussion ends with

[2]Endemol UK ensured us that the four contestants do not know each other before the show and enter and leave the television studio separately, so that they have no opportunity to make any further arrangements after the show.

[3]Players have only limited information about the lottery, i.e., they only know that there may be doubles and they know the margins of the distribution. But they do not know the distribution of the remaining 98 balls.

each contestant secretly casting a vote against one of the other contestants. On the basis of the votes, a contestant is eliminated from the show.[4] After the contestant who has to leave is determined, all contestants open their back row balls and thereby reveal whether they stated the truth or not. The four balls of the leaving contestant are out of the game, while the remaining twelve are carried over to round 2.

Round 2 At the beginning of the second round, two new cash balls are drawn from the lottery and one killer ball is added. These three new balls are mixed with the remaining twelve from round 1, and then shuffled and distributed to the three contestants at random. Again each contestant arranges the closed balls in two rows, i.e., two balls are on the front and three balls are on the back row. As in round 1 the two balls on the front row are opened and the content is common knowledge, while the three balls on the back are private information (see Figure I, right). This time the contestants themselves determine the order of stating the content of their back row balls. Like in the first round, a discussion between the contestants follows and afterwards they secretly choose a contestant they want to vote off. After the contestant to leave has been determined all ball values are revealed, and the five balls of the leaving contestant are out of the game.

Final Round The ten balls from round 2 are carried over to the final round and one last ball, a killer, is added. The maximal amount the contestants can gain is the sum of the highest five cash values among the eleven balls. Before the first phase starts, this amount is announced as the potential jackpot by the show host.
In the *first phase* of the final round the two contestants successively select five of the eleven closed and shuffled balls. These five values build the actual jackpot. If a contestant chooses a killer ball for the jackpot the accumulated amount up to that point is reduced to one-tenth of the original value.
In the *second phase* of the final round the contestants play a prisoner's dilemma in which defection is a weakly dominant strategy (see Figure II).[5] Such a prisoner's dilemma has three pure-strategy Nash equilibria, namely (steal, split), (steal, steal), and (split, steal).[6] The dilemma game is played as follows: each contestant is assigned two balls, one with the word "steal" and one with the word "split" inside. Then both contestants choose one of the balls and open it simultaneously. If both contestants chose the split ball, the jackpot (J) is divided equally between the two

[4]The contestant who receives the highest number of votes has to leave the show. In case of a tie the contestants having received no vote can decide which contestant has to leave. If all contestants received one vote each, contestants openly discuss which contestant has to leave. If contestants do not reach a conclusion, ties are broken arbitrarily.

[5]The show host explains the different outcomes of the game in each episode with the same neutral words (for the exact wording see Appendix A.1).

[6]Two of the resulting Nash equilibria involve one contestant to cooperate. Applying the method of iterated elimination of weakly dominant strategies, however, leaves only the (steal, steal) equilibrium, which should be the only one observed. Thus, each contestant has an incentive to defect, because she is never monetarily worse off when doing so.

contestants. If one contestant chooses steal and the other chooses split, the former gets the whole jackpot and the latter receives nothing. If both chose steal, both get nothing.

Figure II: Prisoner's dilemma game

	split (cooperate)	**steal** (defect)
split (cooperate)	$^1/_2$ jackpot , $^1/_2$ jackpot	0 , jackpot
steal (defect)	jackpot , 0	0 , 0

Note: Defection is a weakly dominant strategy

Immediately before the contestants decide which action to choose, they get roughly 30 to 60 seconds time to discuss. All communication in the game show is free-format.

2.2 Data description

"Golden Balls" was first aired on June, 18th 2007 as a late afternoon (5pm) game show and ended December, 18th 2009.[7] In total, we have records of 222 episodes, divided into four series, with 203 regular and 19 special episodes. Ten special episodes consist of contestants who are on the show for the second time. We exclude these ten episodes from the analysis in order to avoid any bias from repeated interactions. The other nine special episodes comprise of contestants of the same sex. The regular episodes always consist of two women and two men and the contestants are on the show for the first time. All 40 episodes of the first series were filmed prior to the show's television premiere, i.e., all contestants in these episodes had no chance to observe others playing the game. In total we use a data set of 212 episodes (848 contestants), divided into 4 series, where series 1, 2, 3, and 4 consist of $N = 160, 232, 300$, and 156 contestants, respectively.
For all episodes we recorded variables describing the contestant characteristics (age, gender, race, occupation, and place of residence) and the game (all true and stated ball values in rounds 1 and 2, the order of claims in both rounds, votes the contestants received and submitted, the potential and actual jackpot size, communication between contestants before and in the final (handshakes, promises), and the final decision). In addition, we recorded the contestant's action-statement, i.e., before the show starts, the contestants are individually and privately asked to explain which action they intend to play in the final. Table V in the Appendix provides an overview of the data.

[7]The show reached up to 2.2 million people per episode which corresponds to a market share of 21% ("ITV strikes teatime gold", guardian.co.uk, July 3rd, 2007).

3 Analysis of cooperative behavior

We observe an average unilateral cooperation rate of 53.8% and contestants successfully manage to cooperate in even 32.6% (mutual defection in 25.0%) of cases.[8] These rates are considerable keeping in mind that unilateral defectors take home three times as much money as mutual cooperators, £15 693 versus £4 784. The average amount of money left on the table due to mutual defection is £14 426, overall that sums up to £1 558 045 being left on the table.

What parameters affect the positive cooperation rate in the prisoner's dilemma and by what means do cooperators identify each other?
We investigate the individual as well as mutual decision outcome of the contestants when playing the prisoner's dilemma. We make use of the bivariate probit model where the dependent variable y_i is the decision of contestant i either to split ($y_i = 1$) or steal ($y_i = 0$). To analyze team cooperation rates, we make use of the inherent ordering of the mutual decision outcomes, and code the team outcome as equaling 0 if both contestants choose steal ($y_i = 0$), equaling 1 if one contestant chooses steal and the other chooses split ($y_i = 1$), and equaling 2 if both contestants choose split ($y_i = 2$).[9] Throughout, to quantify the influence of the explanatory variables on the predicted probability to split (reach a certain mutual outcome) we report marginal effects. If interactions of two variables are included, we compute interaction effects following the method proposed by Norton et al. (2004) and Mallick (2009). For a detailed specification of the estimation method see Appendix A.3.

The regression analysis is divided into two subsections. Section 3.1 provides preliminary results with respect to player characteristics, observational learning, stake size, and reciprocity. Section 3.2 contains the central analysis dealing with variables of communication comprising of truthful and false promises and handshakes, pre-commitment, and lying in the pre-play.

[8]Note that the distribution of outcomes, in particular the mutual cooperation rate, do not coincide with its population moments. E.g., assuming two randomly chosen contestants play the prisoner's dilemma, underlying an average unilateral cooperation rate of 53.8%, we should observe a mutual cooperation rate of 28.9% (compared to the actual 32.6%). The difference of 3.7 percentage points is significant at the 5% level, i.e., contestants manage to coordinate more frequently than theoretically expected. A χ^2-test for the variance in a normal population rejects the null hypothesis of no difference between the observed sample frequencies and the theoretically expected frequencies at the 5%-level (p=0.025).

[9]Following List (2006) we argue that the contestant's mutual decision outcomes "split-split", "split-steal" and "steal-steal" depend on a single index function and thus have an inherent (natural) ordering fitting the ordered probit model rather than the multinomial one.

Table I: Results from binary probit regressions on unilateral cooperation

	Marginal effects							
	Model (1)		**Model (2)**		**Model (3)**		**Model (4)**	
Player characteristics								
Male	-0.067	(0.051)	-0.057	(0.054)	-0.031	(0.077)	0.040	(0.082)
Age (>40 years)	0.182***	(0.054)	0.184***	(0.056)	0.280***	(0.078)	0.248***	(0.089)
White	0.026	(0.119)	0.015	(0.114)	0.060	(0.163)	0.155	(0.171)
London	-0.020	(0.104)	-0.034	(0.104)	0.290***	(0.110)	0.197	(0.132)
Large city	-0.079	(0.080)	-0.075	(0.082)	-0.193	(0.121)	-0.139	(0.130)
England	-0.273***	(0.068)	-0.273***	(0.071)	-0.305***	(0.090)	-0.304***	(0.092)
Student	-0.028	(0.097)	-0.031	(0.093)	-0.122	(0.132)	-0.137	(0.137)
Pensioner	-0.118	(0.166)	-0.059	(0.165)	-0.159	(0.172)	-0.039	(0.228)
Social job (reputation)	-0.035	(0.083)	-0.045	(0.084)	-0.057	(0.114)	-0.057	(0.120)
Index (social closeness)	0.261	(0.207)	0.286	(0.205)	0.424	(0.274)	0.256	(0.293)
Opp. student	0.056	(0.093)	0.085	(0.094)	0.183	(0.122)	0.184	(0.123)
Opp. pensioner	0.100	(0.157)	0.054	(0.176)	0.130	(0.194)	0.074	(0.178)
Opp. social job	0.029	(0.084)	0.034	(0.084)	0.111	(0.111)	0.148	(0.118)
Team large city	-0.099	(0.102)	-0.104	(0.105)	0.122	(0.159)	0.024	(0.159)
Team small city	-0.168**	(0.069)	-0.165**	(0.071)	-0.178*	(0.108)	-0.218*	(0.114)
Observational learning								
Unexperienced (series 1)	-0.005	(0.070)	-0.018	(0.074)	-0.020	(0.151)	0.073	(0.154)
Experienced (series 4)	0.144*	(0.077)	0.147*	(0.080)	0.162*	(0.095)	0.226**	(0.096)
Stake size								
Log(jackpot)	-0.053***	(0.014)	-0.050***	(0.015)	-0.047**	(0.020)	-0.066***	(0.022)
Log(pot. jackpot)	0.109**	(0.051)	0.134**	(0.055)	0.132*	(0.078)	0.175**	(0.082)
Acc. most money			-0.054	(0.060)	0.001	(0.090)	0.003	(0.088)
Selected higher values in bin/win			0.002	(0.054)	0.113	(0.077)	0.150*	(0.079)
Selected most killers in bin/win			-0.016	(0.058)	-0.002	(0.080)	0.013	(0.088)
Reciprocity								
"Should have left the game"			0.143**	(0.062)	0.047	(0.097)	0.090	(0.101)
Communication								
Started discussion	-0.005	(0.049)	0.007	(0.051)	-0.039	(0.078)	-0.039	(0.081)
Handshakes	-0.185***	(0.069)	-0.187***	(0.070)	-0.323***	(0.099)	-0.225**	(0.095)
Promise	-0.076	(0.110)	-0.093	(0.110)	-0.252	(0.174)	-0.215	(0.177)
Handshakes*promise	0.317***	(0.067)	0.326***	(0.106)	0.559***	(0.163)	0.455**	(0.202)
Pre-commitment								
Statement to split							0.488***	(0.063)
Lying in the pre-play								
Lied about value (round 1)			0.050	(0.068)	0.086	(0.092)	0.108	(0.104)
Lied about killer (round 1)			0.048	(0.070)	0.100	(0.106)	0.139	(0.109)
Opp. lied about value (round 1)			0.096	(0.065)	0.133	(0.088)	0.150	(0.093)
Opp. lied about killer (round 1)			-0.099	(0.069)	-0.019	(0.098)	0.032	(0.103)
Lied about value (round 2)			-0.116	(0.075)	-0.147	(0.103)	-0.126	(0.117)
Lied about killer (round 2)			0.035	(0.071)	0.024	(0.108)	-0.010	(0.118)
Opp. lied about value (round 2)			0.012	(0.075)	-0.100	(0.101)	-0.170	(0.111)
Opp. lied about killer (round 2)			0.048	(0.071)	-0.015	(0.100)	-0.051	(0.110)
Wald $\mathcal{X}^2$	50.68***		61.64***		63.36***		112.05***	
Log-Likelihood	-258.93		-250.71		-127.69		-107.75	
Pseudo R^2	0.11		0.14		0.18		0.31	
Adjusted R^2	0.03		0.02		-0.05		0.08	
N	421		421		226		226	
Number of clusters	212		212		161		161	

Note: binary probit regressions of the decision either to "split" ($y_i = 1$) or "steal" ($y_i = 0$) in the prisoner's dilemma game. Model (1) and model (2) report results using the total sample of $N = 422$ finalists; model (3) and model (4) report results using the subsample of $N = 226$ finalists with an action-statement. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to "split". Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

3.1 Player characteristics, learning, stakes, and reciprocity

All following results are reported in Table I, model (1) and (2), and Table VI in Appendix A.2.

(i) Player characteristics

We control for the impact of various own player, opponent, and team characteristics, such as age, gender, race, place of residence, and occupation on cooperative behavior. On a priori grounds, the relation between these (personal) characteristics and social behavior seems to be rather ambiguous. Deriving clear-cut hypotheses about the influence of these characteristics on the player's propensity to cooperate is therefore impossible. Our results are the following.

Race There is no significant correlation between whites and non-whites and cooperative behavior.

Gender There is also no significant difference between the cooperation rates of men and women. This result is in contrast to e.g., Ortmann and Tichy (1999) who find that females cooperate more frequently than males. However, concerning mutual cooperation we find that a male player is significantly less likely to cooperate with a male opponent, indicating that men are more competitive when facing the same sex.

Age Contestants above 40 years are 18 percentage points more likely to cooperate than contestants below 40 years.[10] In addition, gender conditional on age tends to have an effect on the cooperation rate, i.e., women below the age of 40 cooperate more than men below the age of 40 and vice versa for men and women above the age of 40.

Student or retired We also explore cooperative behavior depending on whether a player herself or her opponent is a student or a pensioner. Yet there are no significant effects.

Place of residence Based on a contestant's place of residence, we construct variables indicating whether the contestant lives in England, London, a small or a big city (see summary statistics in Table V in Appendix A.2). Contestants living in England cooperate significantly less likely than contestants from other parts of Great Britain, i.e., a player who lives in England is roughly 27 percentage points more likely to be a defector. In addition, if both finalists live in England they are 19 percentage points less likely to reach the split-split outcome. If both finalists live in a small city or if both live in a large city, then player i's likelihood to cooperate

[10]Note, the results concerning age should not be attached too much weight since the age categories are merely assessed by personal judgment.

decreases by 11 up to 17 percentage points compared to pairs of contestants who are mixed, i.e., one of them is from a small city while the other is from a large city.

Reputation Since the game is played in front of a large television audience and therefore possibly being watched by friends, family member and colleagues, it might be in a contestant's interest to appear trustworthy, depending on her occupational status. For instance, police officers act as role models for observing the law and behaving correctly, or teachers are responsible for a moral education of children. These people might have an incentive to behave in a fair way, especially when it comes to choosing to cooperate or not in the prisoner's dilemma. But also having a socially responsible job could be a sign for being a cooperator itself, because a cooperative person might select into such a job. The variable "social job" identifies roughly 15% of contestants with a socially responsible occupation, e.g., priests, policemen, firemen, childminders, and teachers. However, the estimation results on a contestant's occupational status exhibit no significant effects for a players likelihood to cooperate.

Index for social closeness The sociological literature argues that the degree of similarity between players has an impact on their social interactions, i.e., people are more likely to form social ties with others who are alike. This tendency of people to relate to similar types is referred to as "homophily".[11] Such motivated, we construct an index for the social closeness between players by accounting for players' age, gender, race, occupational status (social job), and place of residence (England). The index ranges from 0 to 1, weighting each component by one-fifth. Concerning the distribution, we observe 3% of pairs of players to have an index-value of 0.2, 16% to have an index of 0.4, 44% to have an index of 0.6, 33% to have an index of 0.8, and 4% to have an index of 1, i.e., almost 80% of players are relatively socially close to each other having an index value between 0.6 and 0.8. However, we do not find that the social closeness between contestants is a significant determinant of cooperative behavior. This result remains valid for different weights attached to the index's input variables.

(ii) Observational learning

Since the first 40 episodes (series 1) have been filmed before the television premiere of "Golden Balls", contestants of series 1 had no chance to observe other contestants playing the game, i.e., they are unexperienced. In contrast, all later episodes have been filmed after the broadcast of series 1. Especially, contestants of series 4 are most

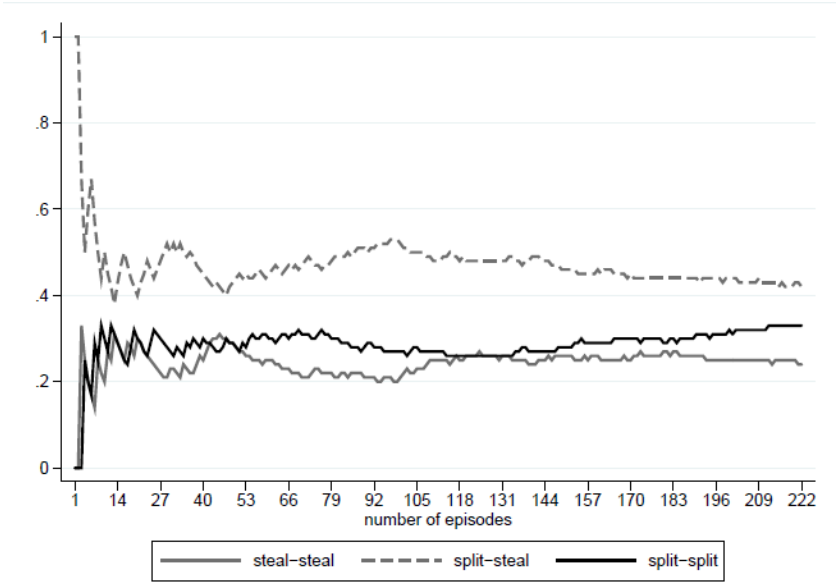
[11] Homophily was first defined by Lazarfeld and Merton (1954), for a survey with respect to sociology see Jackson (2008) and with respect to cooperative game theory see van den Nouweland and Slikker (2001). In many social networks, e.g., friendships or business relations, one observes that individuals associate disproportionately with others who are similar to themselves, usually with regard to gender, race, age, region and education.

familiar with the show and thus might be better in assessing whether cooperation or defection could be successful or not.

Hypothesis 1. *Contestants in later episodes (series 4) use their familiarity with the show and therefore learn whether cooperation could be successful or not.*

Decomposing the data per series reveals that the unilateral cooperation rate is around 50% in series 1-3 and that it jumps to 65.4% in series 4. Most strikingly, successful cooperation even jumps from 30.0% to 48.7% and successful defection declines from 45.0% to 33.3% (for an illustration see Figure III below).

Figure III: The evolution of cooperation over episodes



We observe a substantial difference in the cooperation rate as well as in the distribution of outcomes across series 1 (first 40 episodes) and series 4 (last 39 episodes). A two-sided binomial probability test can reject both, the null hypothesis of no difference between the cooperation rate of experienced (65.4%) and unexperienced (52.5%) players (p=0.000), and the null hypothesis of no difference between the probability of mutual cooperation of experienced (48.7%) and unexperienced (30.0%) players (p=0.000).[12] The estimation results exhibit a positive significant effect of experienced players (series 4) on their likelihood to cooperate. In particular, a player in series 4 is 14 percentage points more likely to cooperate compared to players in series 1-3, and mutual cooperation is 20 percentage points more likely. This suggests that contestants learn over time how to credibly signal their willingness to cooperate, confirming the result of Oberholzer-Gee et al. (2010), who find an effect of

[12]Unless stated otherwise, all tests used in this chapter are two-sided binomial probability tests.

learning for players in later episodes of "Friend or Foe".[13] Summarizing, we cannot reject Hypothesis 1.

(iii) Stake size

The effect of the stake size on cooperation rates in dilemma games is widely debated and no clear answer has been found so far. While some experiments show that an increase in monetary stakes has no significant effect on the cooperation or contribution rate (e.g., Kocher et al., 2008), others suggest that the rate decreases with the stake size (e.g., Camerer and Hogarth, 1999).
Special to the structure of the game show, contestants are confronted with an *actual* and *potential* stake size. At the beginning of the final round the size of the potential jackpot is announced by the show host, which comprises the maximal amount the player can gain. More precisely, the potential jackpot is the sum of the highest five cash values out of the ten balls the two finalists carried over to the final round. Then, in the first phase of the final round the actual jackpot is built by the alternating selection of five balls (see Section 2.1). The size of the potential jackpot varies between £5 000 and £168 100, and the size of the actual jackpot varies between £3 and £100 150.[14]

Actual jackpot We find that the cooperation rate is significantly negatively correlated with the actual jackpot (see Table I and Table VI in Appendix A.2). A 1% increase in the actual jackpot decreases a contestant's likelihood to cooperate by 5 percentage points. But the effect of the actual stakes disappears for stakes above the median. Interestingly, there is a cutoff at £500: the cooperation rate declines sharply from 73.6% for jackpots below £500 to 50.3% for jackpots above £500. In addition, the mutual cooperation rate is significantly higher if the actual jackpot is below £500 (55.6%) than if it is above £500 (27.8%). Both differences are highly significant (p=0.000). The result is even more remarkable if one bears in mind that a stake size around £500 is already much higher than the one used in most laboratory experiments.

Potential jackpot At the same time, however, we find a countervailing effect of stake size: The cooperation rate rises with an increase in the potential jackpot. Contestants are 10-13 percentage points more likely to cooperate the higher the potential jackpot. Hence, the effects of the actual and the potential jackpot operate in opposite directions. But the effect of the potential jackpot on cooperation disappears if we exclude the actual jackpot from the regression. We tested for various

[13]We also separately tested for an effect of series 1, 2, and 3 on the likelihood to cooperate, and find no significant effect.

[14]Stakes in "Golden Balls" are at highest level compared to laboratory experiments as well as related studies using television game show data. These include the US game show "Friend or Foe" analyzed by List (2006) and Oberholzer-Gee et al. (2010) and the Dutch show "Deelt ie't of deelt ie't niet" studied by Belot et al. (2010).

relations between the potential and actual jackpot, determining the ratio and difference between both. The difference has a significant negative impact on cooperation. A 1% increase in the difference increases a player's likelihood to cooperate by 5 percentage points (p=0.056, table unreported). For the ratio between the actual jackpot and the potential jackpot the reverse is true. A player is 20 percentage points less likely to cooperate the larger the ratio, i.e., the closer the spread between actual and potential jackpot (p=0.064, table unreported).

Expectation One might presume that the contestants' perception of the actual jackpot depends on the potential jackpot, i.e., two actual jackpots equal in size will be judged differently depending on their difference to the potential jackpot. Contestants might build an expectation about the size of the actual jackpot depending on the observed size of the potential jackpot. Since computing the mathematically correct expectation is a rather difficult task - especially due to the jackpot-damaging-power[15] of killer balls - contestants need some alternative method to calculate the expectation. As mentioned before, the episodes of series 1 have been broadcasted before all other episodes were filmed. Since contestants in later episodes could observe the average ratio between the jackpot and the potential jackpot in series 1, we assume them using this ratio to form an estimate of the expected jackpot. The average jackpot in series 1 is £13 066, which corresponds to 27.5% of the average potential jackpot of £47 526. This ratio is multiplied by the observed potential jackpot in each episode and determines the contestants' expected jackpot.[16]
We test for a relation between the expectation of the jackpot and cooperative behavior. We find that depending on whether the jackpot is above or below the contestant's expectation, the propensity to cooperate changes. The cooperation rate is significantly higher if the jackpot is below the expectation, and contestants are 16-17 percentage points more likely to cooperate (p=0.002, table unreported). Also cooperation is much less successful if the expectation threshold is taken, i.e., the mutual cooperation rate declines from 41.1% to 18.4%.

Entitlement Experimental research has accumulated evidence that people's preferences over income distribution reflect other-regarding preferences. People care about the distributional consequences of an action, in particular, they make their choice dependent on their individual perception of the others' "worthiness of compensation", see e.g., Rutström and Williams (2000), Hoffman and Spitzer (1985).[17] We test whether contestants in "Golden Balls" might perceive the jackpot allocation

[15]Recall, if a killer ball is chosen for the jackpot the accumulated amount up to that point is reduced to one-tenth of the original value.

[16]In episodes following series 1 the average ratio between the jackpot and the potential jackpot is 25.7%, which is very similar.

[17]Hoffman and Spitzer (1985) examine the effects on individual choices when initial entitlements are allocated at random versus when subjects are required to earn those, and find that the frequency of non-self-interested behavior is lower when initial entitlements are allocated according to pure chance.

differently if they contributed more to the final stake size by either carrying over the higher balls' values from round 2 to the final or by selecting the higher balls' values during the first phase of the final round. Either way, these contestants might feel entitled to a larger "piece of the pie", since they contributed more to both the potential and the actual jackpot and show this by a lower propensity to cooperate. We construct two dummy variables, one for the contestant who contributes most to the potential jackpot, and one for the contestant who selects the highest values when building the actual jackpot, and formulate the following hypothesis.

Hypothesis 2. *A contestant who contributes most to the potential jackpot and/or a contestant who selects the higher values for the actual jackpot is less likely to cooperate.*

Our results show that both having accumulated more money as well as having selected the higher values have a negative but not significant impact on the contestants' likelihood to cooperate. Thus, we find no support for Hypothesis 2.

(iv) Reciprocity

Lastly, we have the opportunity to account for variables linked to kindness, perceived kindness, and its repayment. It is well reported by theoretical and experimental research that people care about underlying motives and intensions of their actions, i.e., whether their behavior is perceived to be fair by one another (see e.g., Rabin, 1993; Falk and Fischbacher, 2006).
Recall that after each pre-play round contestants need to vote off one opponent. However, as we will point out in Section 4, the contestant to leave is not always the one with the lowest ball values, although from a purely monetary perspective, this contestant should be voted off the game. If such a contestant nevertheless makes it to the final, she might respond in her final decision to the kindness she received from her opponent during the pre-play, since (at least in round 2) she owes her "survival" to her opponent's voting decision. We construct a variable labeled "should have left the game" ranking the three contestants in round 2 with respect to their weighted sum of cash values and killer balls. The dummy points at the contestant with the lowest weighted sum.[18]

Hypothesis 3. *A contestant who "should have left the game" is more likely to split the jackpot.*

In the regression analysis we find a significant and positive effect on the contestants likelihood to cooperate: A contestant who should have left the game is roughly

[18]In order to rank the contestants we use the *ex-post* cash-killer-criterion which is described and discussed in detail in Appendix A.4. We assume that a contestant, who does have the lowest weighted monetary amount is aware of this, e.g., often contestants address their pass to the final round during the final discussion and thank their opponent for having taken her so far.

14 percentage points more likely to split which we interpret as reciprocating her opponent's confidence (see Table I model (2)). Hence, we find strong support for Hypothesis 3.[19]

3.2 Verbal and non-verbal communication

We now turn to variables of verbal and non-verbal communication. In particular, we address the impact of truthful and false promises and handshakes, pre-commitment, and lying in the pre-play on cooperative behavior.

(i) Promises and handshakes

Shortly before the prisoner's dilemma is played, the two finalist roughly get 30-60 seconds to discuss what they intend to play, i.e., to cooperate or defect. This communication is free-format. Usually the time is used to ensure one another the honesty of their intention to split, corroborated by a handshake, to thank the other for taking her through the game, or to apologize for lying during the pre-play.
From a theoretical point of view in a prisoner's dilemma, communication is cheap talk that should not affect peoples' behavior (see e.g., Crawford, 1998; Farrell and Rabin, 1996). Various experimental studies, however, have reported that communication has a considerable influence on cooperative behavior, even in one-shot situations, especially if it involves a mutual agreement to cooperate.[20] The effectiveness

[19]We also considered a second channel for reciprocal behavior in line with van den Assem et al. (2012). Consider the possibility that a contestant i makes it to the final, although she received one or two votes during the pre-play. If the final opponent j had cast a vote against i, then j's behavior might be interpreted as unkind by agent i, i.e., the vote expresses j's dislike against i. Therefore i might be willing to punish j for her unkind behavior, i.e., she might be less likely to cooperate. We estimate the same probits of Table I, including a dummy variable indicating whether contestant j voted against contestant i in the pre-play. We indeed find that contestant i is more likely to defect when contestant j had cast a vote against her. However, we exclude the dummy in our main analysis, since including it reduces the data set by 55 observations due to a voting result of 2:1:1:0 in round 1 or 1:1:1 in round 2 (in these cases it is analytically not possible to trace back the contestants' individual voting decision). Except for the episodes with a tie, the outcome of the voting decision is 2:1:0 in round 2, such that none of the final contestants received a vote from their opponent. Hence, the control variable only comprises of the voting result in round 1. Since after round 1 the contestants can only speculate who had cast a vote against whom, we cannot control for whether a contestant received a vote by her opponent in the final, only for the voting decision made by the particular contestant herself. Therefore we can only identify 18 out of 390 contestants to have casted a vote against her final opponent, which leaves the variable with not much explanatory power.

[20]For surveys on the effect of communication see Sally (1995) and Ledyard (1995). In laboratory experiments free-form written communication is often used instead of face-to-face verbal and non-verbal communication to be able to disentangle the effect of facial expressions from the bare content of communication. Roth (1995) provides a survey of bargaining experiments in which he shows that face-to-face communication increases the chance of reaching an agreement even further than free-form messaging. Bohnet and Frey (1999) observe an increase in the unilateral cooperation rate up to 78% if they allow for face-to-face communication.

of communication differs by the words used. In a meta-analysis Sally (1995) estimates that the solicitation of promises by the experimenter raises the cooperation rate in prisoner's dilemma games by 12%-30%. Vanberg (2008) finds supporting evidence that people have a preference for keeping a promise and are not driven by concerns about their expected payoff, and Ellingsen and Johannesson (2004) even propose that people have a preference for keeping their word per se. In contrast, Charness and Dufwenberg (2006) develop the idea that people keep promises because of guilt aversion.[21] Belot et al. (2010) explicitly contrast the effects of voluntary vs. elicited promises in a prisoner's dilemma environment. While elicited promises are uninformative, contestants are roughly 30%-60% more likely to cooperate if they voluntarily promised to share.

Only little work has been done on non-verbal communication and its influence on cooperative behavior so far. For instance, Scharlemann et al. (2001) investigate the impact of a smiling face on people's behavior in a one-shot trust game and find that subjects are significantly more likely to trust when shown a smiling photograph of their counterpart with whom they believe to play. Manzini et al. (2009) address this issue in the minimum effort game and test whether people's propensity to choose high effort is increased if subjects can send a "smile" to the other contestant instead of pressing an ordinary "ready to play" button. They find that this simple device helps contestants to coordinate on a higher effort even though contestants are not able to see or to talk to each other. Psychologists agree that handshakes have a signaling effect, i.e., they convey information about a person's personality and are important for first impressions (see e.g., Chaplin et al., 2000; Stewart et al., 2008). They also have a significant effect in social interactions, e.g., the willingness to help another person increases after having been touched (see Argyle, 1988).

We conjecture that verbal and non-verbal communication serve as an instrument for the contestants to credibly signal their willingness to cooperate. We derive the following hypothesis.

Hypothesis 4. *A contestant who promises to split the jackpot and/or who shakes hands with her opponent to corroborate her intention to split is more likely to cooperate.*

We observe that 25% of pairs of contestants voluntarily promise each other to cooperate. In addition and most interestingly, we observe that 40% of pairs of contestants voluntarily use handshakes to corroborate their intention to split, and 34% out of those pairs of contestants do both, they shake hands and promise each other to share the jackpot. Testing Hypothesis 4, we can verify that verbal and non-verbal communication affect contestants' behavior, but we find a countervailing effect (see Table I, model (1) and (2)). Regarding non-verbal communication, if both finalists shake hands during the final discussion, a contestant is actually 19 percentage

[21] In related work, Miettinen and Suetens (2008) show that contestants feel most guilty if they communicated their intention to cooperate, but then defect while the opponent cooperates.

points more likely to defect than if she does not shake hands, and both finalists are also 12 percentage points less likely to reach the "split-split" outcome (see Table VI in Appendix A.2). This is surprising, since we expected handshakes to serve as a positive commitment device, rather than to serve as an instrument to manipulate the opponent's attitude towards cooperation. Here, when contestants shake hands they lie about their intention. In contrast, if both finalists promise each other to split the jackpot and corroborate their promise via a handshake, each contestant is 32-33 percentage points more likely to cooperate. Further, we only find limited support that a voluntarily stated mere promises has an impact on cooperation.[22] When both finalists promise each other to split the jackpot, they are more likely to reach the "split-steal" outcome (see Table VI in Appendix A.2).

In addition, looking at the development of handshakes and promises across series, we find that the use of both continuously increases over time. While in series 1, only 7.5% of pairs of contestants promise to split and only 15.0% shake hands, in series 4, already 48.7% promise to split and even 64.1% shake hands. Both differences between series are highly significant, p=0.000. Also, contestants seem to learn the proper use of handshakes: In series 1 the decision either to split or steal after a handshake is 50:50, whereas in series 4 64% of contestants choose steal after shaking hands.

Summarizing, the results show that verbal and non-verbal communication conveys information about contestant's intentions. Promises and handshakes are offered and trusted. For the use of handshakes in combination with a promise we find a positive impact on cooperation, but handshakes without a promise result in a negative effect. Hence, we can partly confirm Hypothesis 4.

(ii) False promises and false handshakes

Although we observe that 60% of contestants hold their promises in combination with a handshake, 40% of contestants are liars. As shown in the previous section, a handshake seems to be used as a manipulating device. Table II reports probit estimates on the probability to lie with respect to a promise (model (1)) or a handshake (model (2)) as a function of contestant characteristics and stake size.

Independent of a contestant's own choice of action, promises and/or handshakes are used to initiate cooperative behavior from the opponent. Since we find that the probability to cooperate is affected by stake size, we expect a positive correlation between stake size and the propensity to lie, i.e., to make a false pledge. Indeed, we find a direct effect of stake size on the probability of lying. An increase in the actual jackpot increases the probability to make a false promise by 18 percentage points and to make a false handshake by 8 percentage points. This finding is in contrast

[22] Our finding is in contrast to Belot et al. (2010), however, they also count a statement of intent as a promise. We define promises more narrowly and only identify a promise if the word itself has been used by the contestants.

Table II: Results from binary probit regressions on a false promise or handshake

	Marginal effects			
	Model (1) false promise		**Model (2) false handshake**	
Communication				
Started discussion	0.068	(0.107)	0.023	(0.083)
Handshakes	-0.205	(0.144)		
Promise			-0.315***	(0.094)
Player characteristics				
Male	-0.066	(0.172)	0.177*	(0.097)
Age (> 40 years)	-0.331***	(0.119)	-0.275***	(0.087)
White	-0.241	(0.284)	0.019	(0.209)
England	0.271*	(0.145)	0.374***	(0.134)
Student	0.357**	(0.157)	0.080	(0.159)
Social job (reputation)	0.127	(0.166)	0.077	(0.136)
Index (social closeness)	-0.285	(0.582)	-0.710*	(0.412)
Opp. age (> 40 years)	-0.050	(0.124)	-0.073	(0.096)
Opp. male	0.018	(0.174)	0.197**	(0.097)
Opp. white	0.242*	(0.134)	0.301**	(0.151)
Opp. England	-0.151	(0.244)	-0.076	(0.160)
Opp. student	-0.162	(0.182)	-0.266*	(0.147)
Opp. social job	-0.060	(0.157)	0.074	(0.136)
Stake size				
Log(jackpot)	0.179***	(0.060)	0.077**	(0.031)
Log(pot. jackpot)	-0.451**	(0.189)	-0.216**	(0.096)
Acc. most money	0.057	(0.117)	0.128	(0.096)
Selected higher values in bin/win	-0.104	(0.094)	0.038	(0.082)
Selected most killers in bin/win	-0.050	(0.132)	0.035	(0.095)
Reciprocity				
"Should have left the game"	-0.184	(0.134)	-0.103	(0.125)
Lying in the pre-play				
Lied about value (round 1)	0.120	(0.171)	-0.058	(0.125)
Lied about killer (round 1)	0.137	(0.164)	0.043	(0.133)
Opp. lied about value (round 1)	-0.044	(0.180)	0.029	(0.131)
Opp. lied about killer (round 1)	0.034	(0.146)	0.149	(0.122)
Lied about value (round 2)	0.537***	(0.115)	0.123	(0.128)
Lied about killer (round 2)	0.156	(0.212)	0.144	(0.119)
Opp. lied about value (round 2)	0.259	(0.195)	-0.250*	(0.133)
Opp. lied about killer (round 2)	-0.036	(0.187)	-0.057	(0.127)
Wald χ^2	56.01***		65.22***	
Log-Likelihood	-46.83		-90.89	
Pseudo R^2	0.34		0.22	
Adjusted R^2	-0.09		-0.03	
N	103		169	
Number of clusters	52		85	

Note: Model (1): binary probit regression of the decision either to "make a false promise" ($y_i = 1$) or to "keep a promise" ($y_i = 0$) in the prisoner's dilemma game. Model (2): binary probit regression of the decision either to "make a false handshake" ($y_i = 1$) or to "keep a handshake" ($y_i = 0$) in the prisoner's dilemma game. The marginal effect of the respective explanatory variable determines the effective change of this variable on contestant i's predicted probability to "make a false promise" ("make a false handshake"). Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

to Belot et al. (2010), who find no effect of stake size on the likelihood to hold a promise.
In addition, we find significant differences in the personal characteristics of contestants. Older contestants are 33 percentage points more likely to keep their promise, while English and students tend to be significantly more likely to lie. Facing a non-white opponent results in an increase of 24 percentage points in the contestant's propensity to break her promise. Males turn out to be 18 percentage points more likely to break their handshake than women. Interestingly, if a contestant faces a male opponent, she is now roughly 20 percentage points more likely to break a handshake than if she faces a women. Again, older contestants are less likely to break their pledge, while English are more likely to do so. Facing a non-white opponent leads to a 30 percentage points higher likelihood to break a handshake.

(iii) Pre-Commitment

We have the opportunity to investigate the effect of early commitment. Before the game show starts, contestants are individually asked to make a private statement about their hypothetical action, either split or steal, in case they reach the final round (see Section 2.2). These filmed action-statements are broadcasted to the television audience, but neither to the (other) contestants nor the audience in the television studio.
The action-statements are publicly observed, such that contestants might feel social pressure to live up to it, since they commit oneself to an action.
Various studies in social psychology investigate whether people have a taste for consistency, or examine the consequences of such preferences for (economic) behavior, see e.g., Festinger (1957), or Freedman and Fraser (1966). Falk and Zimmermann (2011) postulate a theoretical model and report evidence from three experiments that people reveal a preference for consistency. They show that people behave consistently to their initial commitment even though they have received additional action-relevant information. Further, they show that people, who are provoked to make a statement about an hypothetical action, live up to their statement when asked to actually choose their action. Here, social pressure is the driving force.
We coded the contestants' action-statements and use them as a best proxy to test whether contestants choose their actual action in line with their statement.

Hypothesis 5. *Contestants who made an unambiguous action-statement before the show act according to this action-statement.*

We observe an unambiguous action-statement by 53.3% (subsample, N=226) of the finalists. The raw data show that 57.5% of finalists state to steal versus 42.5% who state to split before the game show starts. Interestingly, the majority among those who state to split is female (58.3%), while among those who state to steal the majority is male (56.2%). Men state to steal significantly more often than women

(p=0.006). The observed (actual) cooperation rate of all contestants who made an action-statement is 52.2%, which is almost equal to the observed cooperation rate over the whole sample (53.8%).
In Table III we depict the average cooperation rate depending on the contestants' action-statement.

Table III: Relation between action-statement and the cooperation rate

		Outcome	
Action-statement	Total (%)	Steal (%)	Split (%)
Steal	57.52	64.62	35.38
Split	42.48	25.00	75.00
Total	100	47.8	52.2

Table III shows that 75% (64.6%) of those who stated to split (steal) indeed behave consistent with their action-statement. These contestants account for 69% of finalists in the subsample, and for 40% of finalists in the whole sample.
We control for the action-statement in the regression on cooperative behavior (see Table I, model (3) and model (4), and Table VII in Appendix A.2). If a contestant stated to split, she is actually 49 percentage points more likely to split than if she stated to steal. Regarding the mutual outcomes, if both finalists stated to split (steal), their likelihood to jointly cooperate (defect) increases by 35 (25) percentage points compared to the mixed "split-steal" outcome. The magnitude of the results is reflected by the sharp increase in the goodness of fit of both models after controlling for the action-statement. Comparing model (3) and (4) in Table I, we find an increase in the adjusted R^2 of 0.13. The same pattern is shown in the ordered probits, where the adjusted R^2 increases by 0.02 after including the action-statement (the sample is restricted to 117 observations of pairs of contestants, see Table VII in Appendix A.2). Further, note that the effects of all other explanatory variables remain unchanged, which supports the robustness of our findings. Altogether, a substantial fraction of contestants commits to an action and acts consistently.
However, we also observe that 31% of finalists do not follow their initially stated action. They are significantly more likely to switch from a steal-statement to splitting (35.4%) than from a split-statement to stealing (25.0%), $p = 0.008$. These contestants make their final decision contingent on verbal and non-verbal communication and stake size. Also exogenous contestant characteristics seem to be important. The results form probit regressions on the likelihood to switch from the action-statement are reported in Table VIII in Appendix A.2. Throughout, we find a couple of differences: While it does not matter who started the final discussion for a contestant's

likelihood to switch from steal to split (model (2)), it has a strongly significant impact on the decision to switch to steal (model (1)), i.e., a contestant who started the discussion is 23 percentage points more likely to switch to steal. As in Section 3.2(i), handshakes increase a contestant's likelihood to steal, while handshakes in combination with a promise decrease it. For a contestant who stated to steal the effects on the likelihood to switch to split are (almost) vice versa, but differ in magnitude.
Summarizing, we find sufficient support for Hypothesis 5. A substantial fraction of contestants behaves consistently to their action-statement. For contestants who deviate from their stated action, variables of communication turn out to be the driving force.

(iv) Lying in the pre-play

Contestants may dishonestly state the content of their (hidden) back row balls in the first two rounds of pre-play. Whether a contestant has lied or not is revealed by the show host after each of the two rounds. Contestants are then labeled as a liar and may incur a reputation cost for the rest of the game, which can have an impact on the contestants' behavior in the prisoner's dilemma. On the one hand, a liar may be trusted less such that the opponent is more likely to defect. On the other hand a liar may be more likely to defect, since she already suffers a bad reputation which makes her inhibition threshold for choosing to defect is lower than for an honest contestant. We test whether lying in the pre-play is correlated with non-cooperative behavior.

Hypothesis 6. *(a) A contestant is less likely to cooperate if the opponent has lied in the pre-play. (b) A contestant who lied in the pre-play is more likely to defect.*

Analyzing the raw data, we find that 61.3% of the finalists lie in at least one of the two rounds of pre-play. In round 1, 21.9% of finalists lie about a hidden killer ball, and 24.8% overstate a hidden cash value. In round 2, 23.8% of finalists hide a killer, and 18.2% overstate a cash amount. Addressing the mutual outcomes, we observe that in 19.8% of cases two honest contestants face each other in the final, in 42.5% of cases both contestants are liars, and in 37.7% of cases we have an asymmetric pair of finalists, i.e., one of the two contestants has lied before.
The estimation results provide very limited support for an impact of lying on cooperation: If both finalists did not lie about the cash amount in their hidden back row balls, they are almost 12 percentage points more likely to successfully cooperate than if they lied (see Table VI in Appendix A.2). Including lying in the regression on unilateral cooperation, we find no significant effect on cooperative behavior (see Table I, model (2)). Therefore, we find no support for Hypothesis 6. However, we find a strong correlation between lying in the second round of pre-play and lying by

making a false promise to split: A contestant who lied about a cash value is roughly 54 percentage points more likely to also lie in promising to split (see Table II).

4 Lying and partner selection

This section consists of two subsection, where subsection 4.1 addresses the partner selection process during the two rounds of pre-play and subsection 4.2 discusses the role of lying.

4.1 Voting behavior

In each of the two pre-play rounds of Golden Balls the contestants face the decision for whom to vote to leave the show. And in their voting decision the contestant can be strategic in the following sense. Firstly, in line with the show host's prompt "keep in the cash, kick out the killers", contestants may have powerful material (monetary) incentives to cast their vote against the weakest contestant in terms of values or killers in order to maximize stake size. Secondly, in view of their final decision in the prisoner's dilemma, contestants may also want to assess their counterpart's trustworthiness or susceptibility to manipulation. Thirdly, and most importantly, contestants need to ensure their own survival, i.e., they need to make their vote dependent on their belief about the others' voting decision, and need to take into account the cost of lying.
Further, the contestant's voting behavior might vary between the two pre-play rounds. In round 1, material incentives may be attached more weight, since the contestants are "far away" from the final round and neither contestant has been labeled a liar. In round 2, however, contestants may shift weight to personal judgement about the potential opponent's sympathy or trustworthiness with regard to the final. In this sense, the two pre-play rounds give us the opportunity to test whether people strategically vote against certain contestants.

Hypothesis 7. *(a) A contestant who is materially worse off is more likely to be voted to leave the game in round 1 than a contestant who has higher stakes. (b) A contestant who has lied in the previous and/or current round is more likely to be voted to leave the game in round 2 than a contestant who has been honest.*

We investigate the contestants' voting decisions in both rounds using various (ordered) probit regressions on the contestant's propensity to be either voted off the show or to receive a certain number of votes, i.e., 0-3 votes in round 1 and 0-2 votes in round 2. The regression results are reported in Tables IV and IX in Appendix A.2.

Table IV: Results from binary probit regressions on voting behavior

	Marginal effects			
	Round 1		**Round 2**	
Player characteristics				
Male	0.042	(0.036)	0.036	(0.046)
Age (> 40 years)	0.008	(0.035)	0.061	(0.046)
White	-0.132*	(0.069)	-0.116	(0.085)
England	0.025	(0.045)	0.026	(0.051)
Material voting criteria				
Log(total FR)	-0.069***	(0.009)	-0.034***	(0.011)
Log(total claimed BR)	0.014	(0.015)	-0.001	(0.017)
No. killers FR	0.133***	(0.037)	0.046	(0.050)
No. killers claimed BR	0.027	(0.047)	0.070*	(0.041)
Lying				
Lied about value	0.038	(0.036)	0.135***	(0.050)
Lied about killer	0.085**	(0.042)	0.137***	(0.050)
Lied in round 1			0.066*	(0.038)
First to claim	0.022	(0.043)	-0.095**	(0.047)
Wald χ^2	160.66***		59.45***	
Log-Likelihood	-343.05		-353.95	
Pseudo R^2	0.25		0.08	
Adjusted R^2	0.22		0.06	
N	810		607	
Number of clusters	203		203	

Note 1: Binary probit regressions of the probability either to "be voted off in round 1 (2) in the pre-play" ($y_i = 1$) or to "pass round 1 (2) in the pre-play" ($y_i = 0$). The marginal effect of the respective explanatory variable determines the effective change of this variable on contestant i's predicted probability to "be voted off". Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Note 2: FR $\hat{=}$ front row; BR $\hat{=}$ back row.

(i) Material voting criteria

To account for the monetary incentives of the contestants, we construct variables composed of the content of the two front row balls, and the claimed content of the back row balls. Thereby we separately look at the total value (counting killers as zero), and the total number of killers.[23]

In the probit regression on the likelihood to leave the show we find supporting evidence regarding round 1, see Table IV. The larger the number of killers a contestant has or the lower her front row total, the more likely she is voted off the

[23]We also constructed a weighted total of values and killers (where each killer reduces the total to one tenth), but we find no additional effects. The results seem to be driven by the total of values only.

show (increase by 13.3 percentage points and 6.9 percentage points, respectively). The same pattern arises in the results of the ordered probit regressions (see Table IX in Appendix A.2). In round 2, the effect of the number of killers on the front row disappears. Here, contestants take the claimed number of killers seriously. A possible explanation might be that contestants in round 2 have more information about the content of all balls compared to round 1. The only unknown balls in round 2 are the two newly drawn ones from the lottery. For an extensive discussion of the voting behavior with respect to information advantages between contestants and rounds we refer to Appendix A.4. The findings strongly support Hypothesis 7(a).

(ii) Player characteristics

There exists a vast economic as well as psychological literature on racial and gender discrimination.[24] We, however, find only limited support for discriminatory voting. Contestants weakly discriminate against non-whites in round 1 (see Table IV). In round 2 the race effect disappears. Addressing in-group biases, we observe that in both rounds women cast a vote against men significantly more often and vice versa (p=0.000 and p=0.000). In particular, in round 1 (round 2), in 75.4% (66.3%) of cases a woman casts a vote against a man, and in 65.1% (67.0%) of cases a man casts a vote against a woman. Additionally, in round 2 we find a significant difference with respect to voting off the minority gender. Recall, that the remaining contestants in round 2 are either a group of two women and one man or two men and one woman. In a group of two women and one man, we find that the man is voted off significantly more often than the woman in a group of two men and one woman (p=0.082).

(iii) Costs of lying

We control for whether it is a "good" strategy of contestants - in order to survive the pre-play - to lie about the content of their (hidden) back row balls.
We observe that 53% of contestants lie in round 1 and 45% of contestants lie in round 2, and significantly more contestants who have not lied in round 1 reach the second round of pre-play (86.1% vs. 65.27%, p=0.000). Also in the second round liars are less likely to reach the final (54.6% vs. 76.6%, p=0.000). These results suggest that contestants seem to be able to identify and to punish liars.
The regression results support our descriptive findings, see Table IV. In round 2, a contestant is 13.5 (13.7) percentage points more likely to be voted off if she lied

[24]For literature regarding discrimination in the labor market see, e.g., Becker (1957), Turner and Brown (1978), and Altonji and Blank (1999). Levitt (2004) and Anonovics et al. (2005) test taste-based and information-based theories of discrimination, determining whether contestants in the US television game show "The Weakest Link" discriminate on the basis of either gender, age, race, or skill level. Both find patterns consistent with taste-based discrimination with respect to age and gender.

about her total (number of killers). In round 1, contestants only punish concealing a killer: A contestant is 8.5 percentage points more likely to be eliminated from the game if she lied about her number of killers. Lying about a value has no significant effect in round 1. The findings strongly support Hypothesis 7(b). Since contestants have no information about the hidden back row balls of each other, and only know that there have to be four killer balls in total, it is much harder to identify lying with respect to values than to killers in round 1. Therefore, the effects of lying are much more pronounced in round 2. We also control for the effect of a liar of round 1 in the voting decision of round 2, since after round 1 the content of all balls is disclosed and thus lies are uncovered. We find that a contestant is 6.6 percentage points more likely to be voted off in round 2 if she lied in round 1. This is supported by the results from the ordered probit, whether a contestant lied in round 1 significantly matters for the probability of receiving zero or two votes in round 2 (see Table IX in Appendix A.2).
Apart form that, we look at the order in which contestants make their claims each round. Recall, in round 1 the order of claims is exogenously given by the show host, whereas in round 2 the contestants themselves determine the order of claiming the content of their hidden balls (see Section 2). We find that being chosen to claim first has no effect in round 1, but choosing to be the first one in round 2 decreases the likelihood to be voted off by 9.5 percentage points. This suggests that contestants take the claims of the first one more seriously. Indeed, contestants who are first to claim lie significantly less than their followers (40.9% vs. 48.0%, p=0.0419).
Finally, the explanatory power decreases sharply between both rounds. In round 1, a much higher mass of the variance is explained compared to round 2 (see Table IV, Pseudo R^2=0.25 to Pseudo R^2=0.08). This decline serves as an indicator for a switch of contestants' voting incentives between the two pre-play rounds. It is likely that contestants decide on whom to vote off the game by means of sympathy or trustworthiness. Unfortunately, we are not able to directly control for those effects.

To summarize, strategic considerations such as accumulating a high jackpot and selecting a trustworthy counterpart for the final round, are the primary determinants of voting behavior. A contestants personal characteristics only seem to play a minor role. Hence, we find strong support for Hypothesis 7.

4.2 The decision to lie

The effect of lying on economic decision making is highly discussed in the literature. If players have the opportunity to costly punish the other subjects for playing selfishly, Brandts and Charness (2003) show that contestants punish much more often if the selfish action followed a deceptive message. Gneezy (2005), and Fischbacher and Heusi (2008) explore the circumstances under which people lie. Gneezy (2005) uses

a cheap talk sender-receiver game and shows that people's evaluation of whether to lie or not in a particular situation depends on the consequences of the lie in terms of payoffs. Thereby not only gains achievable through lying are considered but also possible losses that might occur to the other contestants. The fraction of liars is largest if the resulting gains are high and the costs, i.e., losses for the other contestants, are low. Contrasting this result, Fischbacher and Heusi (2008) find that the distribution of (partially) truthful, and untruthful people is more or less the same independent of the stake size, the consequences of lying, learning, and the degree of anonymity. It has also been shown that people give themselves away when lying. E.g., Wang et al. (2010) find in a sender-receiver game that peoples' pupils widen significantly when subjects are lying, thereby allowing others to identify a liar. And it is a folk wisdom that people may recognize a liar with the help of certain body signals, for instance, avoiding eye contact, sweating, or blushing.
We define lying as an untruthful claim of the content of the hidden back row balls. While an undetected lie might increase the likelihood to survive in the game, an identified liar might be more likely to be punished by being voted to leave the show in the current or following round. Therefore contestants need to trade off the costs and benefits of lying.
Beyond that, the first round allows us to investigate people's decision to lie and if so, how they lie. We focus on the first round to avoid biases arising out of a player's history from play. Here, contestant would lie in a "rational" way, when they condition their lie on their own (material) position in the game (in terms of open front row balls) relative to her opponents.

Hypothesis 8. *A contestant is more likely to lie if she is a weak player relative to the others in terms of the value of her open front row balls.*

First of all, one could think that observed lies are only "white" lies, i.e., only contestants who are worst off in terms of their balls' content use a lie to increase their odds of survival. Looking at those contestants we find that 77.5% indeed lie.[25] But 28.3% of contestants lie even though they do not have the weakest balls. Following a liar from round 1, we find that she is also more likely to lie in the second round compared to contestants that have not lied in the first round (50.9% vs. 39.9%, p=0.000). As expected a pairwise correlation test shows that lying in the first and second round is positively correlated (ρ=0.109) and highly significant (p=0.006). This result is surprising, because the balls are reshuffled after round 1 and randomly allocated to the contestants in round 2.
We also run various probit regressions on the propensity to lie, as well as ordinary least squares regressions on the difference between the claimed and the true values,

[25] A contestant is defined to be worst off if she has either the lowest total, the highest number of killers, or the lowest weighted total (this corresponds to the prediction generated by the ex-post material voting criteria, see Appendix A.4).

see Table X and Table XI in Appendix A.2, respectively.
The results in Table X show that contestants are significantly more likely to lie about their balls' content if they have one or two killers on their front or back row as well as if they have a low total. These effects are also present if we consider lying about a value (model (2)) and lying about a killer (model (3)) separately. The main difference is that contestants are additionally taking the other contestants' front row into account when making the decision to lie about a killer. A contestant is significantly more likely to lie about a killer if the others have a high total and she is less likely to lie if the others have at least one or two killers. That a contestant seems to neglect the other contestants' position in the game could be interpreted in terms of level-k reasoning. A level-1 contestant makes her decision to lie only dependent on her own position in the game, since she believes that her opponents are level-0. A level-0 contestant uniformly randomizes over all actions (lying, not lying).[26] Level-k behavior is present through all series of the game show. However, an unexperienced contestant is significantly more likely to lie than an experienced contestant. Running probits separately for contestants of the first and later series and comparing the results, we find stronger evidence for level-1 behavior among unexperienced contestants. This suggests that observational learning does not help the contestants to overcome their limited sophistication entirely. In addition, as already pointed out above, the contestant's decision to lie seems to be correlated with the order of precedence in making claims. The contestant who is first to make her claims is 8.8 percentage points less likely to lie about a killer.
Conditional on the fact that a contestant is lying we also analyze by how much a contestant chooses to overstate her balls' content. We find that, on average, a contestant's lie amounts to £6 442. Table XI shows that a contestant's overstatement is significantly higher the higher the other contestant's total, and is significantly lower the better the contestant's own position in the game. Interestingly, males are lying by a significantly lower amount than females.
To summarize, the decision to lie as well as the chosen amount of a lie is driven by a contestant's own position in the game, but the strength and weaknesses of the other contestants are ignored in many respects. We therefore cannot fully support Hypothesis 8. Contestants show limited sophistication and are rather bounded rational in their decision to lie.

5 Conclusion

In this study we investigate the formation of cooperative behavior in a high stakes prisoner's dilemma with face-to-face communication and two stages of pre-play. We

[26]The level-k model was introduced by e.g., Stahl and Wilson (1994).

use data from a British television game show, which consists of two rounds of pre-play and a third round in which two final contestants play a (weak) one-shot prisoner's dilemma game.
The unilateral cooperation rate is 54%, and the mutual cooperation rate is 33%. The main findings are that verbal and non-verbal communication as well as stake size have a powerful influence on cooperative behavior. A promise in combination with a handshake increases the likelihood to cooperate significantly. However, a handshake alone serves as manipulating device in the sense that a contestant's probability to cooperate significantly decreases when she offers a handshake. Further, we find a negative correlation between (expected) stake size and cooperation.
There is also a strong link between the contestants behavior in the pre-play and the behavior in the prisoner's dilemma. A contestant who, from a monetary perspective, should have been voted off the game show in the pre-play, but who nevertheless reaches the prisoner's dilemma game, reciprocates the perceived kindness, i.e., her survival, with an increased propensity to cooperate.
The data also reveal several interesting insights with respect to lying and the perception and consequences of lies: A revealed liar is punished by a higher propensity to be eliminated from the game. With respect to the decision to lie, contestants seem to be bounded rational, i.e., contestants make their lie dependent on their own position in the game, but neglect the strength and weaknesses of their opponents. Finally, we test for consistent behavior of the contestants in the sense that they act accordingly to their pre-announced intended action. The data reveal that a substantial fraction of contestants who explicitly state their intended action for the prisoner's dilemma before the show, actually live up to it.
Our results corroborate the relevance and informational value of communication in a field setting, and augment the existing literature with evidence on non-verbal communication.

Acknowledgements

The authors thank Armin Schmutzler, Nick Netzer, Michelle Sovinsky, Tore Ellingson, Michael Kosfeld, Daniel Schunk, Jacob Goeree, Kevin Staub, Leif Brandes, and seminar participants at Zurich, at the ESA Meeting 2010 in Copenhagen, at the RES Annual Conference 2011 in London, at the 2011 Workshop on the Determinants and Implications of Prosocial Behavior in Southampton, at the Verein für Socialpolitik Annual Meeting 2011 in Frankfurt, and at the Zurich Workshop on Economics in Lucerne 2011 for helpful discussions and suggestions. Financial support of the Swiss National Science Foundation is gratefully acknowledged. The data were provided to the authors by the television show producers, courtesy of Endemol UK plc, in May 2009.

Appendix

A.1 Instructions of the prisoner's dilemma

The show host Jasper Carrott explains the "weak" prisoner's dilemma in every episode with almost the same words:

> "It is time to split or steal. You have got two final golden balls left, you have each got a golden ball with the word *split* written inside, you have got each a golden ball with the word *steal* written inside. I will ask you to make a conscious choice and you will choose either the split or the steal ball, neither of you will know what the other has chosen. If you both choose the split balls, you split today's jackpot of $£J$ and you both go home with $£J/2$. If one of you splits and one of you steals, whoever steals goes home with all the money $£J$, whoever splits goes home with nothing. If you both decide to steal and you are very greedy, you both go home with nothing. Before I ask you to choose, Player A, B just check the two balls to make sure you know which is to split and which is to steal. Do not show to each other. It is very important that you know which is which. [PLAYERS CHECK THE BALLS] Are you happy to know which is split and which is steal? Okay, before I ask you to choose, I will give you some time to talk to each other about what has happened today and how you feel. [PLAYERS DISCUSS] Okay, player A, B choose the split or steal ball now. [PLAYERS CHOOSE BALLS] Hold it up, make sure that when you open it, the other player can see it. Player A, B split or steal? [PLAYERS OPEN BALLS]"

A.2 Tables

Table V: Summary statistics

Variable	Mean	Std. dev.	Min	Max	N
Series (1 = series 1, 2 = series 2, 3 = series 3, 4 = series 4)	2.56	0.50	0	4	888
Player characteristics					
Social job[a] (1 = social job)	0.14	0.34	0	1	887
Student (1 = student)	0.08	0.27	0	1	888
Pensioner (1 = retired)	0.03	0.17	0	1	888
England (1 = England, 0 = SCO, WAL, NIR, IRL)	0.85	0.36	0	1	886
Large City[b] (1 = population > 268 300)	0.30	0.46	0	1	886
London (1 = London)	0.13	0.34	0	1	888
Gender (1 = male)	0.50	0.50	0	1	888
Race (1 = white)	0.92	0.27	0	1	888
Age[c] (1 = above 40 years)	0.43	0.50	0	1	888
Action-statement[d] (0 = steal, 1 = split, 2 = other)	1.08	0.86	0	2	612
Average cash ball in the show	5619.55	10374.12	10	75000	3108
Round 1					
Value of open balls (balls 1 and 2)[e]	8802.64	13858.91	0	104000	888
Value claimed for balls 3 and 4	14265.86	13908.53	0	83000	888
Value of closed balls (balls 3 and 4)	7852.88	12315.07	0	83000	888
Number of killers in open balls	0.47	0.58	0	2	888
Number of killers claimed	0.23	0.43	0	2	888
Number of killers in closed balls	0.53	0.60	0	2	888
Player lied at least about one ball	0.53	0.50	0	1	888
Player lied at least about one value	0.32	0.47	0	1	888
Player lied at least about one killer	0.28	0.45	0	1	888
Number of killers taken to round 2	2.59	0.76	1	4	888
Round 2					
Value of open balls (balls 5 and 6)	9651.32	14275.73	0	103000	666
Value claimed for balls 7, 8 and 9	18421.19	16683.73	105	95000	666
Value of closed balls (balls 7, 8 and 9)	13352.47	16291.90	0	95000	666
Number of killers in open balls	0.44	0.58	0	2	666
Number of killers claimed	0.44	0.52	0	2	666
Number of killers in closed balls	0.75	0.69	0	3	666
Player lied at least about one ball	0.45	0.50	0	1	666
Player lied at least about one value	0.23	0.42	0	1	666
Player lied at least about one killer	0.28	0.45	0	1	666
Number of killers taken to final round	2.14	0.91	0	5	666
Value of balls taken to final round	23003.79	21134.80	150	143300	666
Final round (1st phase)					
Potential jackpot	51238.36	31261.51	5000	168100	444
Average cash ball	6932.27	12030.86	10	75000	1122
Number of killers	3.21	0.94	1	6	144
Number of killers to bin	1.74	0.92	0	4	144
Number of killers to win	1.47	0.88	0	4	144
Final round (2nd phase)					
Jackpot	13343.03	19247.56	3	100150	444
Ratio (jackpot/potential jackpot)	0.25	0.28	0.0001	1	444
Decision (1 = split)	0.54	0.50	0	1	444
Money taken home	4916.96	12000.86	0	100150	444
Money taken home (steal / split)	15693.11	20087.90	3	100150	94
Money taken home (split / split)	4783.64	8440.02	1.83	43950	148
Money left on the table	14426.34	20255.76	100	92330	108
Discussion (1 = starts discussion)	0.5	0.5	0	1	444
Handshake (1 = shake hands)	0.39	0.49	0	1	444
Mutual promise (1 = say promise)	0.25	0.43	0	1	444
Handshake*promise (1 = shake hands and say promise)	0.16	0.37	0	1	444

[a] A social job is defined as a job in which people care for other people, e.g., doctors, nurses, child minders, social workers, teachers, police officers, firemen, soldiers.

[b] Large cities are cities with more than 268,300 inhabitants (based on the Mid-2008 Population Estimates published by the Office for National Statistics).

[c] Contestants are estimated to be below or above the age of 40 by personal judgment.

[d] Contestants secretly state their intended action for the final before the show starts. Note, that the action-statement is not filmed in the first 18 episodes of series 1. This reduces the data set to 194 episodes.

[e] Killer balls are counted as zero for all value variables.

Table VI: Results from ordered probit regressions on outcomes in the PD

	Marginal effects					
	steal/steal (0)		**split/steal (1)**		**split/split (2)**	
Player characteristics						
Team male	0.159	(0.108)	-0.007	(0.029)	-0.151*	(0.083)
Team female	0.009	(0.073)	0.002	(0.012)	-0.010	(0.085)
Team > 40 years	-0.080	(0.071)	-0.029	(0.041)	0.109	(0.110)
Team < 40 years	0.128*	(0.076)	0.011	(0.014)	-0.139*	(0.076)
Team England	0.140**	(0.055)	0.050	(0.032)	-0.190**	(0.081)
Team small city	0.118**	(0.055)	0.022	(0.017)	-0.140**	(0.065)
Team large city	0.168*	(0.101)	-0.016	(0.033)	-0.152**	(0.072)
Index (social closeness)	-0.342	(0.230)	-0.065	(0.057)	0.407	(0.272)
Learning						
Unexperienced (series 1)	-0.024	(0.064)	-0.005	(0.017)	0.029	(0.081)
Experienced (series 4)	-0.134**	(0.056)	-0.062	(0.051)	0.196*	(0.104)
Communication						
Handshakes	0.103*	(0.061)	0.020*	(0.011)	-0.122**	(0.062)
Promise	-0.044	(0.066)	0.112***	(0.017)	0.053	(0.082)
Handshakes*promise	-0.306***	(0.108)	0.034	(0.100)	0.340***	(0.067)
Stake size						
Log(jackpot)	0.048***	(0.013)	0.009	(0.006)	-0.058***	(0.015)
Log(pot. jackpot)	-0.092**	(0.045)	-0.017	(0.015)	0.110**	(0.055)
Lying round 1 and 2						
Team lied cash	-0.089	(0.062)	-0.033	(0.040)	0.122	(0.100)
Team lied killer	0.019	(0.070)	0.003	(0.010)	-0.022	(0.079)
Team never lied Cash	-0.096*	(0.053)	-0.021	(0.018)	0.117*	(0.067)
Team never lied Killer	0.022	(0.057)	0.004	(0.010)	-0.025	(0.067)
Wald $\mathcal{X}^2$	42.23***					
Log-Likelihood	-203.42					
Pseudo R^2	0.11					
Adjusted R^2	0.02					
N	212					
Number of clusters	212					

Note: Ordered probit regressions on the mutual decision outcomes in the prisoner's dilemma game, where the mutual outcome is coded as equaling 0 if both contestants choose "steal", as equaling 1 if one contestant chooses "steal" and the other contestant chooses "split", and as equaling 2 if both contestants choose "split". The marginal effect determines how a change in the respective explanatory variable changes the distribution of the outcome. The "team variables" are indicators and equal 1 if the team is so composed and 0 otherwise, e.g., "team female" equals 1 if both contestants are female, and 0 otherwise. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table VII: Results from ordered probit regressions on outcomes in the PD, subsample of contestants who made an unambiguous action-statement

	Marginal effects					
	steal/steal (0)		**split/steal (1)**		**split/split (2)**	
Player characteristics						
Team male	0.125	(0.131)	-0.007	(0.032)	-0.118	(0.105)
Team female	0.074	(0.128)	0.001	(0.016)	-0.076	(0.117)
Team > 40 years	-0.082	(0.114)	-0.026	(0.063)	0.108	(0.175)
Team < 40 years	0.168	(0.108)	-0.003	(0.027)	-0.165	(0.095)
Team England	0.101	(0.081)	0.024	(0.031)	-0.125	(0.106)
Team small city	0.083	(0.079)	0.011	(0.020)	-0.095	(0.094)
Team large city	0.077	(0.126)	-0.002	(0.019)	-0.075	(0.110)
Index (social closeness)	-0.436	(0.350)	-0.055	(0.078)	0.490	(0.388)
Learning						
Unexperienced (series 1)	0.010	(0.131)	0.001	(0.011)	-0.011	(0.142)
Experienced (series 4)	-0.121	(0.083)	-0.040	(0.056)	0.161	(0.134)
Communication						
Team handshakes	0.132	(0.100)	0.008	(0.027)	-0.140*	(0.082)
Team promise	-0.015	(0.100)	-0.063***	(0.008)	-0.016	(0.108)
Team handshakes*promise	-0.561***	(0.193)	-0.023	(0.197)	0.584***	(0.093)
Stake size						
Log(jackpot)	0.060***	(0.018)	0.007	(0.009)	-0.067***	(0.020)
Log(pot. jackpot)	-0.116*	(0.060)	-0.014	(0.020)	0.130*	(0.068)
Lying round 1 and 2						
Team lied cash	-0.050	(0.114)	-0.012	(0.043)	0.062	(0.156)
Team lied killer	0.027	(0.101)	0.002	(0.005)	-0.029	(0.105)
Team never lied cash	-0.069	(0.075)	-0.010	(0.019)	0.080	(0.090)
Team never lied killer	-0.013	(0.078)	-0.002	(0.011)	0.014	(0.088)
Commitment						
Team statement split	-0.190***	(0.054)	-0.160*	(0.085)	0.350***	(0.125)
Team statement steal	0.251**	(0.108)	-0.044	(0.051)	-0.208***	(0.070)
Wald $\mathcal{X}^2$	53.14***					
Log-Likelihood	-103.70					
Pseudo R^2	0.18					
Adjusted R^2	-0.02					
N	117					
Number of clusters	117					

Note: Ordered probit regressions on the mutual decision outcomes in the prisoner's dilemma game, where the mutual outcome is coded as equaling 0 if both contestants choose "steal", as equaling 1 if one contestant chooses "steal" and the other contestant chooses "split", and as equaling 2 if both contestants choose "split". Here, subsample of 53.3% of contestants who make an unambiguous action-statement (either "split" or "steal") before the show. The marginal effect determines how a change in the respective explanatory variable changes the distribution of the outcome. The "team variables" are indicators and equal 1 if the team is so composed and 0 otherwise, e.g., "team female" equals 1 if both contestants are female, and 0 otherwise. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table VIII: Results from binary probit regressions on switching from the action-statement

	Marginal effects			
	Model (1) switch to steal		**Model (2) switch to split**	
Player characteristics				
Male	-0.065	(0.067)	-0.004	(0.099)
Age (> 40 years)	-0.227***	(0.082)	0.210*	(0.109)
White	-0.282	(0.210)	-0.023	(0.243)
London	-0.133***	(0.047)	0.006	(0.174)
England	0.121**	(0.054)	-0.280**	(0.142)
Student	-0.047	(0.119)	-0.142	(0.144)
Pensioner	0.361	(0.337)	-0.189	(0.159)
Social job (Reputation)	0.014	(0.130)	0.050	(0.142)
Index (social closeness)	-0.112	(0.224)	0.709*	(0.365)
Opponent characteristics				
Opp. student	-0.156***	(0.044)	0.112	(0.150)
Opp. pensioner	-0.115*	(0.061)	-0.192	(0.170)
Opp. social job	0.060	(0.137)	0.358***	(0.137)
Communication				
Started discussion	0.226***	(0.083)	0.071	(0.093)
Handshakes	0.253**	(0.149)	-0.253***	(0.096)
Promise	0.435*	(0.222)	-0.193	(0.191)
Handshake*promise	-0.689***	(0.162)	0.505***	(0.112)
Stake Size				
Log(jackpot)	0.089***	(0.029)	-0.054**	(0.023)
Log(pot. jackpot)	-0.131*	(0.078)	0.248***	(0.095)
Wald χ^2	32.18**		29.77**	
Log-Likelihood	-37.09		-68.97	
Pseudo R^2	0.31		0.18	
Adjusted R^2	-0.04		-0.04	
N	96		130	
Number of clusters	87		108	

Note: Model (1): binary probit regression of the decision of player i either to choose "steal, when announced to split in the action-statement" ($y_i = 1$) or to "split, when announced to split in the action statement" ($y_i = 0$) in the prisoner's dilemma game. Model (2): binary probit regression of the decision of player i either to "split, when announced to steal in the action-statement" ($y_i = 1$) or to "steal, when announced to steal in the action-statement" ($y_i = 0$) in the prisoner's dilemma game. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to "steal, when announced to split in the action-statement" ("split, when announced to steal in the action-statement"). Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table IX: Results from ordered probit regressions on the number of votes

	Marginal effects									
	No. of votes = 0		No. of votes = 1		No. of votes = 2		No. of votes = 3			
Round 1										
Player characteristics										
Male	-0.050	(0.036)	0.007	(0.005)	0.025	(0.018)	0.019	(0.013)		
Age (> 40 years)	0.008	(0.035)	-0.001	(0.005)	-0.004	(0.017)	-0.003	(0.013)		
White	0.144**	(0.059)	-0.005	(0.006)	-0.070**	(0.028)	-0.070	(0.037)		
England	-0.026	(0.046)	0.004	(0.008)	0.013	(0.023)	0.009	(0.016)		
Material voting criteria										
Log(total FR)	0.097***	(0.010)	-0.013***	(0.002)	-0.048***	(0.006)	-0.036***	(0.005)		
Log(total claimed BR)	-0.043	(0.016)	0.006**	(0.002)	0.021***	(0.008)	0.016***	(0.006)		
No. killers FR	-0.141	(0.040)	0.019***	(0.007)	0.069***	(0.020)	0.052***	(0.015)	Wald χ^2	290.01***
No. killer claimed BR	0.023	(0.048)	-0.003	(0.007)	-0.011	(0.024)	-0.009	(0.018)	Log-Likelihood	-834.70
Lying									Pseudo R^2	0.19
Lied about value	-0.103***	(0.036)	0.011***	(0.004)	0.051***	(0.018)	0.042***	(0.016)	Adjusted R^2	0.18
Lied about killer	-0.147***	(0.038)	0.012***	(0.003)	0.072***	(0.019)	0.063***	(0.019)	N	810
First to claim	0.028	(0.044)	-0.004	(0.007)	-0.014	(0.021)	-0.010	(0.015)	Number of clusters	203
Round 2										
Player characteristics										
Male	-0.010	(0.038)	0.000	(0.000)	0.010	(0.038)				
Age (> 40 years)	-0.036	(0.036)	-0.000	(0.001)	0.036	(0.037)				
White	0.033	(0.077)	0.001	(0.007)	-0.034	(0.083)				
England	-0.032	(0.042)	0.001	(0.003)	0.031	(0.040)				
Material voting criteria										
Log(total FR)	0.040***	(0.010)	-0.000	(0.001)	-0.040***	(0.010)				
Log(total claimed BR)	-0.002	(0.015)	-0.000	(0.000)	-0.002	(0.015)				
No. killers FR	-0.047	(0.041)	0.000	(0.001)	0.047	(0.041)				
No. killer claimed BR	-0.045	(0.034)	-0.000	(0.001)	0.045	(0.034)			Wald χ^2	93.30***
Lying									Log-Likelihood	-613.93
Lied about value	-0.100**	(0.039)	-0.008	(0.007)	0.108**	(0.046)			Pseudo R^2	0.08
Lied about killer	-0.143***	(0.036)	-0.013*	(0.008)	0.156***	(0.043)			Adjusted R^2	0.06
First to claim	0.160***	(0.046)	-0.011	(0.007)	-0.148***	(0.040)			N	607
Lied in Round 1	-0.071**	(0.031)	-0.000	(0.007)	0.072**	(0.032)			Number of clusters	203

Note 1: Ordered probit regressions on the contestant's likelihood to receive a certain number of votes in each pre-play round, per episode. The marginal effects determine how a change in the respective explanatory variable changes the distribution of the outcome. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).
Note 2: FR $\hat{=}$ front row; BR $\hat{=}$ for back row. 11 special episodes, in which all contestants have the same sex are excluded.

Table X: Results from binary probit regressions on lying in round 1

	Marginal effects					
	Model (1) Lied		Model (2) Lied about value		Model (3) Lied about killer	
Player characteristics						
Male	-0.033	(0.044)	-0.053	(0.033)	0.061*	(0.033)
Age (> 40 years)	0.058	(0.046)	0.067*	(0.038)	0.060*	(0.036)
White	-0.008	(0.074)	-0.069	(0.071)	-0.051	(0.061)
London	-0.246***	(0.080)	-0.113*	(0.060)	0.056	(0.063)
Large city	0.137***	(0.049)	0.041	(0.040)	0.011	(0.042)
England	0.004	(0.062)	-0.014	(0.048)	-0.033	(0.050)
Student	-0.101	(0.077)	-0.075	(0.060)	0.037	(0.063)
Pensioner	0.130	(0.105)	0.212**	(0.102)	-0.119*	(0.062)
Social job (reputation)	0.009	(0.059)	-0.034	(0.044)	0.026	(0.049)
Unexperienced (series 1)	0.210***	(0.055)	0.119***	(0.045)	0.099**	(0.039)
Experienced (series 4)	-0.099*	(0.054)	-0.003	(0.043)	-0.096***	(0.029)
First to claim	0.004	(0.052)	0.002	(0.041)	-0.088**	(0.036)
Last to claim	-0.020	(0.049)	-0.008	(0.041)	-0.022	(0.040)
Balls' content						
Log(total own FR)	-0.092***	(0.013)	-0.074***	(0.009)	-0.000	(0.006)
Log(total own BR)	-0.114***	(0.011)	-0.041***	(0.007)	-0.080***	(0.007)
Log(total oth. FR)	0.049*	(0.026)	-0.015	(0.019)	0.046***	(0.016)
1,2 killers own FR	0.384***	(0.039)	0.106***	(0.037)	0.209***	(0.036)
1,2 killers own BR	0.374***	(0.040)	-0.158***	(0.039)		
1,2 killers oth. FR	-0.055	(0.060)	0.054	(0.047)	-0.178***	(0.047)
3,4 killers oth. FR	-0.106	(0.089)	-0.037	(0.069)	-0.234***	(0.026)
Wald χ^2	337.3628***		187.0794***		238.8875***	
Log-Likelihood	-350.02		-432.25		-368.86	
Pseudo R^2	0.40		0.18		0.26	
Adjusted R^2	0.37		0.15		0.22	
N	845		845		845	
Number of clusters	212		212		212	

Note 1: Model (1): binary probit regression of the decision of player i to "lie about a killer and/or value" ($y_i = 1$) or to "be honest about her balls' content" ($y_i = 0$) in round 1. Model (2): binary probit regression of the decision of player i to "lie about her balls' cash values" ($y_i = 1$) or to "not to lie about her balls' cash values" ($y_i = 0$) in round 1. Model (3): binary probit regression of the decision of player i to "lie about her killers" ($y_i = 1$) or to "be honest about her killers" ($y_i = 0$) in round 1. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to "lie". Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Note 2: FR $\hat{=}$ front row; BR $\hat{=}$ back row.

Table XI: Results from ordinary least squares regressions on the difference between the stated and the true amount in round 1

	Model (1)		Model (2)		Model (3)	
	Lied		Lied about value		Lied about killer	
Player characteristics						
Male	-0.276**	(0.128)	-0.258	(0.158)	-0.213	(0.195)
Age (> 40 years)	0.204	(0.138)	0.314*	(0.164)	0.011	(0.209)
White	-0.032	(0.286)	-0.300	(0.297)	0.092	(0.407)
London	0.111	(0.284)	-0.086	(0.344)	0.291	(0.415)
Large city	-0.028	(0.170)	0.151	(0.210)	-0.263	(0.276)
England	-0.139	(0.183)	-0.362*	(0.197)	-0.173	(0.292)
Student	-0.014	(0.267)	0.155	(0.346)	-0.095	(0.339)
Pensioner	-0.292	(0.438)	-0.297	(0.458)	-0.321	(0.584)
Social job (reputation)	-0.217	(0.224)	-0.227	(0.247)	-0.277	(0.359)
Unexperienced (series 1)	0.632***	(0.154)	0.716***	(0.171)	0.463**	(0.200)
Experienced (series 4)	-0.057	(0.215)	0.030	(0.190)	-0.314	(0.343)
First to claim	0.060	(0.171)	-0.195	(0.218)	0.495*	(0.254)
Last to claim	0.213	(0.153)	0.167	(0.188)	0.140	(0.221)
Balls' content						
Log(total own FR)	-0.202***	(0.027)	-0.156***	(0.032)	-0.238***	(0.046)
Log(total own BR)	-0.151***	(0.025)	-0.071	(0.058)	-0.162***	(0.028)
Log(total oth. FR)	0.277***	(0.083)	0.215**	(0.094)	0.326***	(0.120)
1,2 killers own FR	0.187	(0.159)	0.462**	(0.210)	0.157	(0.202)
1,2 killers own BR	-0.029	(0.154)	0.542***	(0.191)		
1,2 killers oth. FR	0.001	(0.162)	-0.144	(0.196)	0.074	(0.232)
3,4 killers oth. FR	-0.071	(0.294)	-0.059	(0.337)	-0.352	(0.483)
F-Statistic	6.64***		4.78***		4.69***	
R^2	0.23		0.26		0.28	
Adjusted R^2	0.20		0.20		0.21	
N	448		270		233	
Number of clusters	209		172		173	

Note 1: Ordinary least squares regressions on the difference between the claimed and true balls' contents (y_i) for subsamples of contestants who "lied" (model (1)), "lied about a cash value" (model (2)), and "lied about a killer" (model (3)) in round 1. The regression coefficients can be interpreted as the change in the expected value of y_i associated with a one-unit increase in a control variable, holding all other control variables constant. Standard errors are reported in parentheses and are corrected for episode clusters. * (p¡0.10), ** (p¡0.05), *** (p¡0.01).
Note 2: FR ≙ front row; BR ≙ back row.

A.3 Estimation method

A.3.1 Modeling the individual decision

In order to explore the individual decision process when playing the prisoner's dilemma game, we make use of the bivariate probit model (see e.g., Wooldridge, 2010, Chap. 15). We build a (index) model around the latent (non-linear) regression

$$y_i^* = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i, \quad \varepsilon_i \sim NID(0, \sigma^2) \tag{A.1}$$

where y_i^* is unobservable, α is a constant, $x_1, ..., x_k$ are (demographic) player characteristics as well as (endogenous) variables evolved during the game, β denotes the response coefficient vector, and ε_i is the random error that is normally, independently, and identically distributed (NID). We only observe the sign of y^* that determines the value of the observed binary response y_i, that is

$$y_i = 0 \quad \text{if} \quad y_i^* \leq 0; \quad y_i = 1 \quad \text{if} \quad y_i^* > 0.$$

We can now compute the conditional probability that player i chooses split as

$$Pr(y_i = 1|\mathbf{X_i}) = Pr(y_i^* > 0|\mathbf{X_i}) = \Phi(\alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} = \Phi(m),$$

where $\alpha, \beta_1, ..., \beta_k$ are parameters to be estimated, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and m denotes the index $\alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}$.
In order to quantify the influence of the explanatory variables on the predicted probability to split we calculate marginal effects. The marginal effect for the j-th independent (continuous) variable is computed as

$$ME_j = \frac{\partial Pr(y_i = 1|\mathbf{X_i})}{\partial x_{ji}} = \phi(m)\beta_j \quad j = 2, ..., K$$

where $\phi(\cdot)$ is the standard normal density function. The magnitude of the derivative is proportional to $\phi(m)\beta_j$. If x_j is a dummy variable, the marginal effect is computed as the discrete difference

$$ME_j^{\text{discrete}} = \frac{\triangle Pr(y_i = 1|\mathbf{X_i})}{\triangle x_{ji}} = \Phi(m|x_{ji} = 1) - \Phi(m|x_{ji} = 0), \quad j = 2, ..., K.$$

Hence, the marginal effect of an explanatory variable is the effect of an effective percentage change of this variable on player i's predicted probability to split, given that all other explanatory variable are held constant.
We also include various interactions of two discrete (dummy) variables in the regression. Here, the interaction effect is defined as the change in player i's predicted probability to split for a change in both interacted variables. For the estimation of the interaction effect we follow the method proposed by Norton et al. (2004). In particular, for the two

interacted dummy variables x_1 and x_2, we model the conditional probability to split as

$$Pr(y_i = 1|\mathbf{X_i}) = \Phi(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + ...\beta_k x_{ki}),$$

with $\alpha, \beta_1, \beta_2, \beta_{12}, ..., \beta_k$ being the parameters to be estimated. The interaction effect is equal to the discrete double difference

$$\begin{aligned} ME_{12} =& \frac{\triangle^2 Pr(y_i = 1|\mathbf{X_i})}{\triangle x_{1i} \triangle x_{2i}} \\ =& \Phi(m) - \Phi(\beta_1 + \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki} \\ & - \Phi(\beta_2 + \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}) + \Phi(\alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}). \end{aligned}$$

Overall, we compute marginal effects at the sample mean. We judge the goodness of fit by the adjusted McFaddens Pseudo R^2, that is a measure for the proximity of the model to the observed data. Standard errors are computed by the delta method. The test statistic follows a χ^2-distribution with degrees of freedom equal to the number of restrictions.

A.3.2 Modeling the mutual decision

In order to analyze team cooperation rates, we estimate ordered probits (see e.g., Wooldridge, 2010, Chap. 15). Here, we make use of the inherent order of outcomes by coding the team outcome as equaling 0 if both players choose steal ($y_i = 0$), equaling 1 if one player chooses steal and the other chooses split ($y_i = 1$), and equaling 2 if both players choose split ($y_i = 2$). The model for the latent variable is identical to the one of the bivariate probit (see equation (A.1)). The relation between the observed variable y_i and the latent variable y_i^* is given by

$$y_i = 0 \quad \text{if} \quad y_i^* < \tau_1; \quad y_i = 1 \quad \text{if} \quad \tau_1 \leq y_i^* < \tau_2; \quad y_i = 2 \quad \text{if} \quad y_i^* \geq \tau_2.$$

The boundaries between the three cases are determined by the thresholds τ_1 and τ_2, which need to be estimated along with the rest of the parameters. The probabilities of the three events $y_i = 0; 1; 2$ are given by $Pr(y_i = 0) = \Phi(\tau_1 - m)$, $Pr(y_i = 1) = \Phi(\tau_2 - m) - \Phi(\tau_1 - m)$, $Pr(y_i = 2) = \Phi(m - \tau_2)$, with $m = \alpha + \beta_1 x_{1i} + ... + \beta_k x_{ki}$.
Since we are interested how a change in an explanatory variable changes the distribution of the outcome variable, we compute marginal effects. In particular, the marginal effect of a variable x_j for the $l-$th response is given by

$$ME_{jl} = \frac{\partial Pr(y_i = l|\mathbf{X_i})}{\partial x_{ji}} = [\phi(\tau_{l-1} - m) - \phi(\tau_l - m)]\beta_j,$$

where again ϕ denotes the standard normal density function. If x_j is a dummy variable, the marginal effect is computed as the discrete difference

$$ME_{jl}^{\text{discrete}} = \triangle Pr(y_i = l|\mathbf{X_i}) = Pr(y_i = l|(\mathbf{X_i}) + \triangle x_j) - Pr(y_i = l|(\mathbf{X_i})).$$

The marginal effect of an interaction between two discrete variables will again be treated differently, since the marginal effects of the interacted variables involve the coefficient of the interaction term, i.e., $m = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + ... + \beta_k x_{ki}$. In the computation we follow the method proposed by Mallick (2009), i.e., the marginal effect of x_1 (analog for x_2) for the l-th response is calculated as

$$ME_{1l} = \frac{\partial Pr(y_i = l|\mathbf{X_i})}{\partial x_{1i}} = [\phi(\tau_{l-1} - m) - \phi(\tau_l - m)](\beta_1 + \beta_{12} x_{2i});$$

and the magnitude of the interaction effect for the l-th response is computed as

$$\begin{aligned} ME_{12l} = \frac{\partial Pr(y_i = l|\mathbf{X_i})}{\partial x_{1i} \partial x_{2i}} =& [\phi(\tau_{l-1} - m) - \phi(\tau_l - m)]\beta_{12} \\ & - [\phi'(\tau_{l-1} - m) - \phi'(\tau_l - m)](\beta_1 + \beta_{12} x_{2i})(\beta_2 + \beta_{12i} x_{1i}), \end{aligned}$$

where $\phi'(\cdot)$ denotes the first derivative of the normal density function w.r.t. its argument. As above, standard errors are computed by the delta method.

A.4 Voting behavior

A.4.1 Material voting criteria

Assuming that it is a contestant's aim to maximize stake size, the contestant's voting decisions can be described by means of material voting criteria. For each episode and round we define three criteria, each pointing at the contestant who is in the weakest position in material terms with respect to the particular criterion.

- **Cash-Criterion (CC)**: constructed on the basis of the monetary values of balls and declares the contestant with the lowest amount of money to be voted to leave the game (a killer ball is counted as a ball with zero value).
- **Killer-Criterion (KC)**: accounts for the number of killer balls per contestant and declares the contestant with the highest number of killers to be voted off the game.
- **Cash-Killer-Criterion (CKC)**: constructed on the basis of the weighted monetary values of balls and declares the contestant with the lowest amount of money to be voted to leave the game (a killer ball reduces the total to one-tenth of the original value, a second killer ball reduces it to one-hundredth and so forth).

Within each criterion we distinguish three different time-dimensions:

- **Ex-ante**: refers to the content of the two balls on the front row.
- **Claimed**: refers to the content of the two front row balls and the claims about the hidden back row balls.
- **Ex-post**: refers to the content of all revealed balls.

By means of these three, respectively nine criteria, we analyze to what extent each criterion explains the contestant's voting decision within and between round 1 and 2.

Descriptive results

Descriptive results are reported in Table XII-XIV in Appendix A.4.3. Note that we exclude all episodes which have a voting result of 2:1:1:0 in round 1 or a tie in round 2, since it is analytically impossible to reconstruct the contestants' individual decision for these episodes. Additionally, in Table XIII and Table XIV in Appendix A.4.3 we restrict the sample to only those contestants who take part in both rounds (with an almost equal share of males (48.5%) and females (51.5%)); thereby we can compare the voting results for both rounds using the decisions of the same contestants.
First we want to look at the proportions of contestants who are effectively voted off in line with the three criteria (Table XII in Appendix A.4.3). Focusing on the time-dimensions of each criterion, in round 1 we find that most contestants vote in line with the prediction of the *ex-ante* CKC and *ex-ante* CC, as well as the *claimed* KC. In round 2 instead, the *ex-post* CKC and the *ex-post* CC dominate, but the *claimed* KC again yields the best prediction. Overall the KC, especially when looking at the *claimed* values, fits best: In round 1, 81.1% of contestants who are voted off have the highest number of killer balls both on their front row as well as claimed on their back row; in round 2 this proportion slightly reduces to 70.3%, but still exceeds the CC and CKC.
The findings are confirmed when looking at the proportions of contestants who receive a vote when predicted by each criterion. We additionally distinguish the data by gender (Table XIII in Appendix A.4.3). As above, focussing on the time-dimensions of each criterion, we find that contestants most frequently vote in line with the *ex-ante* CKC and *ex-ante* CC in round 1, but in line with the *ex-post* CKC and *ex-post* CC in round 2. Concerning the KC, contestants vote in line with the *claimed* KC in round 1 and the *ex-ante* KC in round 2. Separating these findings by gender, we observe that significantly more males than females vote in line with the *claimed* KC and *ex-ante* CKC in round 1 ($p = 0.024$ and $p = 0.034$) whereas in round 2 more females vote in line with the *ex-post* CC ($p = 0.021$).
Most noticeable, in both rounds the contestants take the claims about killer balls seriously, although one might argue that claims are only cheap talk and should therefore be ignored. But a claim about a killer is "self-signaling": If a contestant states to have a killer ball it is the truth.
Concluding, contestants seem to base their voting decision on the *ex-ante* criteria in round 1, and put more weight on the *claimed* and *ex-post* criteria in round 2. This switch maybe explained by the different extent of information a contestant has at a certain time.

A.4.2 Information based separability of contestants

Players have different information about the distribution of balls' values depending on the round of pre-play. In round 1, the contestants face a situation in which they must base their voting decision on an *ambiguous* distribution of outcomes, i.e., only the two values of the front row balls are common knowledge. In round 2, the contestants have additional information. After round 1, all contestants have to reveal the content of their back row balls. All twelve balls that are then carried over from round 1 to round 2 are now common knowledge to the three remaining contestants, and only the two newly added cash values

are unknown. Therefore, a contestant can be in two different states: First, if the two new values are within the revealed balls on the front rows, or if a particular contestant has at least one of the two new values on her own back row and the other is observable on any other contestant's front row, she knows the exact distribution of values in play. From an informational point of view a contestant who knows all ball values in play makes her voting decision in a situation where only the precise allocation of the values is *uncertain*. Second, if both new values remain unobservable, a contestant again lacks information, but not as much as in round 1.[27] Naturally the extent of information should influence the contestants' voting decision, i.e., in round 2 contestants are able to infer - up to a certain extent - whether a contestant claims the truth. We expect that a contestant, who faces *uncertainty* about the allocation of balls - she is able to infer the true content of balls in play - to base her voting decision on the *ex-post* criteria. Instead, we expect contestants who face *ambiguity* about the distribution of balls' values - she cannot infer the true content of all balls in play - to vote most frequently by means of the *ex-ante* criteria.
In Table XIV in Appendix A.4.3 we present proportions of the contestants' voting decision by means of the three criteria and their time-dimensions as well as contestants' informational background. We had to restrict the data set to 573 contestants in order to compare the same contestants in round 1 and 2. In round 2, we identify 50 contestants to be in an *uncertain* state, and 241 contestants to be in an *ambiguous* state. In round 1, all 573 contestants are in the same *ambiguous* situation.
As suggested, we find that the proportion of *uncertain* contestants who vote in line with the *ex-post* CKC is significantly higher than the one of *ambiguous* contestants (p=0.004 and p=0.000). Additionally, the spread between proportions of contestants who vote in line with the *ex-post* or the *claimed* CKC is much larger for contestants facing *uncertainty* than *ambiguity*. For instance, in round 2 70% of all contestants facing *uncertainty* vote by means of the *claimed* CKC, but only 55% of all contestants in round 1, compared to 64% of contestants facing *ambiguity* in round 2. The same holds for the *ex-post* CKC, as well as for males and females with respect to both criteria. Besides we find a different voting pattern for males and females facing *uncertainty*: In round 2, a much larger proportion of males votes in line with the *ex-ante* CC and *ex-ante* CKC than females. But the proportions are almost equal when contestants are either *ambiguous* or in round 1.

Summary With the help of the nine (material) criteria we find a possibility to explain the voting decision of more than two thirds of contestants.[28] Player give most weight to the three killer criteria, and, in round 2, they seem to be able to infer the truth behind the claims.

[27]We also analyze the case when only one of the two new values is known to a contestant, termed *partial ambiguity* (see Table XIV in Appendix A.4.3). But for the sake of clarity we limit the discussion to the cases when both new values are either known or not.

[28]Over the whole sample, we find only 22 contestants (2.8%) who do not vote in line with neither criterion, and only 65 contestants (8.3%) who never vote in line with an *ex-ante* criterion.

A.4.3 Tables voting behavior

Table XII: Voting decision by means of objective criteria

	After round 1[b]		**After round 2**	
Criteria to predict a player	**in**	**out**	**in**	**out**
who should be voted to leave[a]	Row %	Row %	Row %	Row %
Cash-Criterion (CC)				
ex-ante				
stay	87.2	12.8	72.3	27.7
vote to leave	38.3	61.7	55.4	44.6
claimed				
stay	76.7	23.3	70.0	30.0
vote to leave	69.8	30.2	59.9	40.1
ex-post				
stay	83.0	17.0	74.8	25.2
vote to leave	50.9	49.1	50.5	49.5
Killer-Criterion (KC)				
ex-ante				
stay	92.0	8.0	83.6	16.4
vote to leave	23.9	76.1	32.9	67.1
claimed				
stay	93.7	6.3	85.1	14.9
vote to leave	18.9	81.1	29.7	70.3
ex-post				
stay	86.6	13.4	80.9	19.1
vote to leave	40.1	59.9	38.3	61.7
Cash-Killer-Criterion (CKC)				
ex-ante				
stay	87.7	12.3	73.2	26.8
vote to leave	36.9	63.1	53.6	46.4
claimed				
stay	83.3	16.7	74.1	25.9
vote to leave	50.0	50.0	51.8	48.2
ex-post				
stay	83.2	16.8	74.5	25.5
vote to leave	50.5	49.5	50.9	49.1
N	666	222	444	222

[a] We take into account that the prediction might not be unique per episode, i.e., more than one player might have a prediction to be eliminated.

[b] Each round, 222 contestants are eliminated. In round 1 (2) 55.4% (52.2%) of the eliminated players are men.

Table XIII: Voting decision per player by means of objective criteria

Players voted by means of	**all (%)**	**men (%)**	**women (%)**
Cash-Criterion (CC)			
ex-ante			
Round 1	72.4	74.5	70.5
Round 2	66.5	67.3	65.8
claimed			
Round 1	36.0	38.5	33.6
Round 2	64.4	62.6	66.1
ex-post			
Round 1	55.3	55.4	55.3
Round 2	70.2	66.9	73.2
Killer-Criterion (KC)			
ex-ante			
Round 1	82.1	83.0	81.3
Round 2	81.9	84.9	79.1
claimed			
Round 1	83.6	85.5	81.8
Round 2	78.9	80.2	77.6
ex-post			
Round 1	70.9	72.7	69.2
Round 2	76.3	77.0	75.6
Cash-Killer-Criterion (CKC)			
ex-ante			
Round 1	71.0	74.1	68.1
Round 2	65.8	67.3	64.4
claimed			
Round 1	55.0	58.3	51.9
Round 2	63.5	61.9	65.1
ex-post			
Round 1	60.4	61.5	59.3
Round 2	66.8	65.5	68.1
N[a]	573	278 (48.5%)	295 (51.5%)

[a] Note: For purpose of comparability, we restrain the sample to 573 observations including only those players, who are not being eliminated in round 1. Further we consider only those decisions for which we can trace back for whom a player voted, i.e., we exclude episodes with a voting of 2:1:1:0 in round 1 and 1:1:1 in round 2.

Table XIV: Voting decision per player under risk or ambiguity (by means of objective criteria)

	Round 2			Round 1
Players vote by means of	**Uncertainty** (%)	**Partial ambiguity** (%)	**Ambiguity** (%)	**Ambiguity** (%)
Cash-Criterion (CC)				
ex-ante				
all	64.0	66.3	67.2	72.4
men	82.4	64.9	67.7	74.5
women	54.5	67.6	66.7	70.5
claimed				
all	58.0	67.7	61.8	36.0
men	52.9	65.7	60.6	38.5
women	60.6	69.6	63.2	33.6
ex-post				
all	74.0	70.2	69.3	55.3
men	70.6	66.4	66.9	55.4
women	75.8	73.6	71.9	55.3
Killer-Criterion (KC)				
ex-ante				
all	84.3	79.8	84.0	82.1
men	82.4	81.3	89.1	83.0
women	85.3	78.4	78.3	81.3
claimed				
all	84.0	75.9	81.3	83.6
men	94.1	76.9	81.9	85.5
women	78.8	75.0	80.7	81.8
ex-post				
all	78.0	75.2	77.2	70.9
men	88.2	76.9	75.6	72.7
women	72.7	73.6	78.9	69.2
Cash-Killer-Criterion (CKC)				
ex-ante				
all	60.0	64.9	68.0	71.0
men	76.5	62.7	70.9	74.1
women	51.5	66.9	64.9	68.1
claimed				
all	70.0	61.7	64.3	55.0
men	70.6	59.0	63.8	58.3
women	69.7	64.2	64.9	51.9
ex-post				
all	76.0	66.3	65.6	60.4
men	88.2	64.9	63.0	61.5
women	69.7	67.6	68.4	59.3
N[a]	50	282	241	573

[a] Note: The sample is restricted to the same 573 players in round 1 and 2.

References

Altonji, J., and R. Blank (1999): "Race and Gender in the Labor Market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 3, pp. 3143–3159. New York: Elsevier, 1 edn.

Anonovics, K., P. Arcidiacono, and R. Walsh (2005): "Games and Discrimination: Lessons From The Weakest Link," *Journal of Human Resources*, XL (4), 918–947.

Argyle, M. (1988): *Bodily Communication*. New York: Methuen, 2 edn.

Becker, G. (1957): *The Economics of Discrimination*. University of Chicago Press.

Belot, M., V. Bhaskar, and J. van de Ven (2010): "Promises and Cooperation: Evidence from a TV Game Show," *Journal of Economic Behavior and Organization*, 73(3), 396–405.

Bohnet, I., and B. S. Frey (1999): "The Sound of Silence in Prisoner's Dilemma and Dictator Games," *Journal of Economic Behavior and Organization*, 38(1), 43–57.

Brandts, J., and G. Charness (2003): "Truth or Consequences: An Experiment," *Management Science*, 49(1), 116–130.

Camerer, C. F., and R. M. Hogarth (1999): "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 19 (1-3), 7–42.

Chaplin, W. F., J. B. Phillips, J. D. Brown, N. R. Clanton, and J. L. Stein (2000): "Handshaking, Gender, Personality, and First Impressions," *Journal of Personality and Social Psychology*, 79(1), 110–117.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74(6), 1579–1601.

Crawford, V. (1998): "A Survey on Experiments on Communication via Cheap Talk," *Journal of Economic Theory*, 78, 286–298.

Ellingsen, T., and M. Johannesson (2004): "Promises, Threats and Fairness," *The Economic Journal*, 114(495), 397–420.

Falk, A., and U. Fischbacher (2006): "A Theory of Reciprocity," *Games and Economic Behavior*, 54(2), 293–315.

Falk, A., and F. Zimmermann (2011): "Preferences for Consitency," *Working Paper*.

Farrell, J., and M. Rabin (1996): "Cheap Talk," *Journal of Economic Perspectives*, 10 (3), 103–118.

Festinger, L. (1957): *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Fischbacher, U., and F. Heusi (2008): "Lies in Disguise. An Experimental Study on Cheating," *Thurgau Institute of Economics TWI Working Paper 40*.

Freedman, J., and S. Fraser (1966): "Compliance Without Pressure: the Food in the Door Technique," *Journal of Personality and Social Psychology*, 4(2), 195–202.

Gneezy, U. (2005): "Deception: The Role of Consequences," *American Economic Review*, 95 (1), 384–394.

Hoffman, E., and M. L. Spitzer (1985): "Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice," *Journal of Legal Studies*, 14(2), 259–297.

Jackson, M. (2008): *Social and Economic Networks*. Princeton, Princeton University Press.

Kocher, M. G., P. Martinsson, and M. Visser (2008): "Does stake size matter for cooperation and punishment?," *Economic Letters*, 99(3), 508–511.

Lazarfeld, P., and R. K. Merton (1954): "Friendship as a Social Process: A Substantive and Methodological Analysis," in *Freedom and Control in Modern Society*, ed. by M. Berger, T. T. Abel, and C. C. Page. New York: Van Nostrand.

Ledyard, J. O. (1995): "Public Goods: A Survey of Experimental Research," in *The Handbook of Experimental Economics*, ed. by J. H. Kagel, and A. E. Roth, The Handbook of Experimental Economics, pp. 111–194. Princeton University Press.

Levitt, S. (2004): "Testing Theories of Discrimination: Evidence From Weakest Link," *Journal of Law and Economics*, XLVII, 431–452.

List, J. A. (2006): "Friend or Foe? A Natural Experiment of the Prisoner's Dilemma," *Review of Economics and Statistics*, 88(3), 463–471.

Mallick, D. (2009): "Marginal and Interaction Effects in Ordered Response Models," *MPRA paper No 9617*.

Manzini, P., A. Sadrieh, and N. J. Vriend (2009): "On Smiles, Winks and Handshakes as Coordination Devices," *The Economic Journal*, 119(537), 826–854.

Miettinen, T., and S. Suetens (2008): "Communication and Guilt in a Prisoner's Dilemma," *Journal of Conflict Resolution*, 52(6), 945–960.

Norton, E. C., H. Wang, and C. Ai (2004): "Computing Interaction Effects and Standard Errors in Logit and Probit Models," *The Stata Journal*, 4(2), 154–167.

Oberholzer-Gee, F., J. Waldfogel, and M. W. White (2010): "Friend or Foe? Cooperation and Learning in High-Stakes Games," *Review of Economics and Statistics*, 92(1), 179–187.

Ortmann, A., and L. K. Tichy (1999): "Gender differences in the laboratory: evidence from prisoner's dilemma games," *Journal of Economic Behavior and Organization*, 39, 327–339.

Rabin, M. (1993): "Incorporating Fairness into Games Theory and Economics," *American Economic Review*, 83, 1281–1302.

Roth, A. E. (1995): "Bargaining Experiments," in *The Handbook of Experimental Economics*, ed. by J. H. Kagel, and A. E. Roth, chap. 4, pp. 253–348. Princeton University Press.

Rutström, E. E., and M. B. Williams (2000): "Entitlements and Fairness: An Experimental Study of Distributive Preferences," *Journal of Economic Behavior and Organization*, 43, 75–89.

Sally, D. (1995): "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958 to 1992," *Rationality and Society*, 7(1), 58–92.

Scharlemann, J. P. W., C. C. Eckel, A. Kacelnik, and R. K. Wilson (2001): "The Value of a Smile: Game Theory with a Human Face," *Journal of Economic Psychology*, 22, 617–640.

Stahl, D. O., and P. W. Wilson (1994): "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior and Organization*, 25(3), 309–327.

Stewart, G. L., S. L. Dustin, M. R. Barrick, and T. C. Darnold (2008): "Exploring the Handshake in Employment Interviews," *Journal of Applied Psychology*, 93(5), 1139–1146.

Turner, J., and R. Brown (1978): "Social Status, Cognitive Alternatives and Intergroup Relations," in *Differences Between Social Groups*, ed. by H. Tajfel, pp. 101–140. New York: Academic Press.

van den Assem, M., D. van Dolder, and R. Thaler (2012): "Split or Steal? Cooperative Behavior When the Stakes are Large," *Management Science*, 58(1), 2–20.

van den Nouweland, A., and M. Slikker (2001): *Social and Economic Networks in Cooperative Game Theory*. Berlin: Springer.

Vanberg, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica*, 76(6), 1467–1480.

Wang, J. T., M. Spezio, and C. F. Camerer (2010): "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review*, 100(3), 984–1007.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edn.

Chapter 3

Facing a Dilemma: Cooperative Behavior and Beauty

joint with Donja Darai

"All that glitters is not gold;
Often have you heard that told."

William Shakespeare, Merchant of Venice (1596-1598)

1 Introduction

Beginning with the seminal paper of Biddle and Hamermesh (1998), which identifies a wage gap based on physical attractiveness using labor market data of the U.S. and Canada, economists started to be interested in the effects of physical attractiveness on economic decision making. For instance, Mobius and Rosenblat (2006) show that the beauty premium for attractive people is present in a controlled laboratory labor market experiment. Eckel and Petrie (2011) provide evidence that people have a willingness to pay for the possibility to see a picture of their counterpart before making their decision in a trust game, suggesting that valuable information is conveyed by the physical appearance of people. However, the transmission channel of the effect of physical attractiveness on economic decision making is still only partly understood. The most prominent theory is "beauty-is-good" stereotyping. People assign a wide range of positive personality traits and abilities to physically attractive people and believe that physically attractive people are, for instance, smarter,

more trustworthy, more likable, and/or more talented (see e.g., Dion et al., 1972). Accordingly, effects of attractiveness are driven by peoples' beliefs about the attractive, which are then reflected in their behavior towards the attractive. But, *not all that glitters is gold*. There is only very limited evidence that the "beauty-is-good" stereotype is accurate, that is, physically attractive people do not behave differently than less attractive people.

This study investigates the relationship between attractiveness and cooperative behavior and shows that the beauty premium is also present in a high-stakes field setting with two-sided communication and previous interaction between players. We provide evidence that physical attractiveness affects cooperative behavior in mixed-gender interactions only. That is, people's behavior is biased towards cooperation when facing someone attractive of the other sex.

Specifically, to study cooperative behavior, we survey data from 211 episodes of the British television game show "Golden Ball". At the end of the show, two finalists play for a jackpot. The rules of this game follow a slightly modified version of a simultaneous one-shot prisoner's dilemma game: if both contestants choose to cooperate, the jackpot is split equally; if one chooses to defect while the other cooperates, the former receives the entire jackpot and the latter gets nothing; if both defect, they both go home empty-handed. The accumulation of the jackpot takes place in two rounds of pre-play previous to the prisoner's dilemma and requires neither the contestants' effort nor cognitive ability. On average, the stake size amounts to £12 912. During the two rounds of pre-play, the two finalists are selected out of four initial contestants. The game show data are then combined with data on the contestants' facial appearance. We asked 728 independent evaluators to rate portrait photographs of the contestants along various dimension such as physical attractiveness, sympathy, trustworthiness, or likeability.

Our analysis shows that contestants are significantly more cooperative towards a facially attractive opponent than towards a less facially attractive opponent, independent of demographic characteristics, as for instance gender and age, stake size, or communication. Facially attractive contestants are therefore awarded with significantly higher earnings in the prisoner's dilemma. Quantifying this beauty premium, a one-standard deviation increase in facial attractiveness, at the margin, causes the contestant's expected earnings to rise by up to 2 153, as well as the contestant's probability to obtain positive earnings by up to 5.9 percentage points. This effect is not driven by non-cooperative behavior of the attractive. With minor qualifications for younger and female contestants, we find no correlation between a contestant's own facial attractiveness and cooperation.

Although the attractiveness effect likewise applies for men and women, we show that it is limited to mixed-gender interactions only. Contestants are only biased in their decision to cooperate when facing an attractive opponent of the other sex. Thus, the attractiveness effect is not present in same-gender interactions. This finding offers

a new perspective on the underlying mechanism of the effects caused by attractiveness. Our results are not driven by people believing that attractive others are more likely to cooperate, but rather by people having a preference to cooperate more with someone towards whom they are personally attracted. Furthermore, we cannot identify a beauty premium or plainness penalty during the pre-play, which clearly supports our conjecture that personal attraction serves as the underlying transmission channel of physical attractiveness on people's behavior in pairwise interactions and that attractiveness matters most when objective information is missing.

The remainder of this study is structured as follows. In section 2 we discuss the related literature. Section 3 provides the hypotheses. Section 4 describes the data sets. Section 5 presents the results. In Section 6 we discuss the results with respect to the underlying transmission channels. Finally, Section 7 concludes.

2 Literature

There is long history of research on physical attractiveness in psychology, sociology, and evolutionary biology, which attracted economists and also political scientists to investigate the impact of physical attractiveness on economic and political outcomes. In this section we provide an overview of the most important results and theories.

2.1 Physical attractiveness and economic success

Several studies show that attractiveness leads to higher economic gains in various economic interactions in the field.

Labor markets Attractive people generally fare better in the labor market, i.e., they are more likely to be hired, earn higher salaries, and are more likely to be promoted. In their seminal work Biddle and Hamermesh (1998) use a broad household survey of the U.S. and Canada and show that physically attractive employees earn about 10 to 15 percent higher salaries than unattractive workers, independent of the occupation. The size of the beauty premium is comparable to the race and gender wage gaps in the U.S. labor market.[1] Mobius and Rosenblat (2006) show that the results even survive in an experimental labor market involving a real-effort task. In the experiment participants in the role of employers pay wages to participants in the role of workers who perform a maze-solving task. Although attractive employees are not better in solving the task than unattractive employees, attractive

[1]Studies on gender wage differences predict men to earn up to 25 percent more than women in the U.S., unadjusted for gender differences in characteristics. But the gender wage gap can be explained to large parts by differences in skills, experience, or occupational types (see O'Neill, 2003).

employees are awarded higher wages, suggesting that employers hold stereotypical expectations about the performance of physically attractive employees. Ruffle and Shtudiner (2010) explore the value of beauty in the hiring process by examining the response rates to CVs sent to companies in Israel. Two CVs that only differ in containing a portrait photograph of the job applicant or not, are sent to the same company. They find that attractive males are significantly more likely to be called back and invited for a job interview than no-picture males and more than twice as likely as plain-looking males. Surprisingly, no-picture females have the highest response rate, which is 22 percent higher than the response rate of plain females and 30 percent higher than the one of attractive females. The gender specific beauty penalty with respect to women contradicts most findings in the literature.

Political elections In political science, several studies show that physical attractiveness has a significant impact on the evaluation of candidates and thereby on electoral outcomes (see the survey by Ottati and Deiger, 2002). For instance, Berggren et al. (2012) study political elections in Finland and show that a one-standard deviation increase in beauty raises the average non-incumbent candidate's votes by 20 percent. Further evidence is provided by Rosar et al. (2008) who show that more attractive candidates receive a significantly higher poll in a German state election, especially when the average attractiveness of the candidates in the electoral district is low. Antonakis and Dalgas (2009) claim that politicians are not elected on the basis of their ability, but that voters rather infer competence form the politicians' facial appearance. Recruiting adults and children raters in Switzerland, they find that naive voters, such as children, can predict results of the 2002 French parliamentary election retrospectively by rating the competence of the candidates from their photographs.

Pro-social behavior Beauty has been analyzed in various experimental settings with strategic interaction, providing evidence that people behave more generously or cooperatively towards physically attractive people. Solnick and Schweitzer (1999) investigate how recipients' physical attractiveness affects offers in an ultimatum game and show that attractive respondents receive significantly higher offers than unattractive respondents. Supporting these results, Rosenblat (2008) shows that female allocators treat physically and vocally attractive recipients more generously in dictator games.[2] Andreoni and Petrie (2008) analyze a repeated public goods game and find that attractive players earn more than unattractive ones. This observed beauty premium is unrelated to the own contributions of attractive players, but can be attributed to the presence of an attractive group member, that increases the other players' contributions to the public good. In the trust game, Wilson and

[2]Rosenblat's finding that physically attractive recipients do not achieve better outcomes in speech-only conditions suggests that voice matters only when combined with an attractive photograph.

Eckel (2006) demonstrate that attractive trustees are trusted more and that others expect attractive players to be more trusting than unattractive players. In addition, the failure to meet these expectations may lead to a "beauty penalty".[3] Stirrat and Perrett (2010) focus on the width-to-height ratio of male faces, which has been shown to be a "sexually dimorphic, testosterone-linked trait" (p. 349) that serves as a predictor for male aggression. In a laboratory experiment on the trust game they find that males with a large width-to-height ratio are trusted less, and are also less trustworthy themselves compared to males with a more narrow facial width. Their finding provides further evidence that trustworthiness is strongly linked and based in facial features. Mulford et al. (1998) examine repeated exchange relations associated with a prisoner's dilemma game. The results indicate that subjects are more cooperative if pooled with an attractive partner, but attractiveness is based on the perception of the one subject that makes the choice.[4]

2.2 Transmission channels of physical attractiveness

There are several theories about the underlying transmission channel of the effect of physical attractiveness on economic success.

"Beauty-is-good" stereotype Ample research in sociology and psychology has established that people attribute a variety of positive characteristics and qualities to physically attractive people, and negative ones to physically unattractive people. In their pioneering study, Dion et al. (1972) claim that "what is beautiful is good" by demonstrating that attractive people are believed to have better career prospects, to possess socially desirable traits, to lead happier lives and to be happier overall. The link between beauty and goodness suggests the existence of a beauty stereotype. This paper spawned a large literature on the physical attractiveness stereotype, showing for both sexes a robust association between physically attractiveness and cognitive ability, competence, sociability, popularity, dominance, sexual experience, mentally health, and social skills (see reviews by Eagly et al., 1991; Feingold, 1992; Langlois et al., 2000).[5] These believes let people treat attractive ones more favorably, resulting in higher economic gains for the attractive. The physical attractiveness stereotype is particularly strong if there is no other information available about a person. This favoritism would be rational, if the "beauty-is-good" stereotype is

[3] In a further experimental study on the trust game, Eckel and Petrie (2011) investigate the informational value of a photograph and the differential desire to acquire this information. Subjects are willing to pay to see the photograph of their partner whom they transact with, indicating that a face has a positive informational value.

[4] It is questionable whether the used attractiveness rating can be taken as an objective measure of the subject's attractiveness because it is not elicited from third-party judges.

[5] Also, research on babyface-appearances indicates that people with a babyface will be - at all ages - attributed childlike traits, i.e., they are perceived as more naive and more honest than people with a more mature facial appearance (see e.g., Berry and McArthur, 1985).

accurate. But there is only very limited evidence. Jackson et al. (1995) show in their meta-analysis that there is no significant correlation between physical attractiveness and intelligence for adults and only a modest one for children. Zebrowitz et al. (2002), and Kanazawa (2011) support this finding, showing that attractiveness is positively correlated with IQ scores from childhood trough middle adulthood, using large samples from the U.S. and UK. Mueller and Mazur (1997) use data from a cohort of military officers and find that recruits with a high ranked facial appearance are also high ranked in their physical fitness. Concerning social skillfulness and likability there is empirical evidence that physically attractive people indeed possess better social skills and are more likable (see e.g., Goldman and Lewis, 1977; Erwin and Calev, 1984).[6] Hope and Mindell (1994) show that the physical attractiveness stereotype is especially accurate for socially skillful and attractive people. Lastly, concerning non-material gains, Anderson et al. (2001) find that physically attractive people receive higher social status in groups.

"Vocal attractiveness" stereotype and negotiation Zuckerman and Driver (1989) and Zuckerman et al. (1990) have shown that physical attractiveness is positively correlated with vocal attractiveness, proclaiming the existence of a vocal attractiveness stereotype, i.e., "what sounds beautiful is good". Rosenblat (2008) develops the theory that the underlying transmission channel between physical attractiveness and economic success is the "negotiation channel". She shows that vocal attractiveness helps to succeed in bargaining interactions and leads to higher economic gains using a laboratory experiment on the dictator game.

Taste-based discrimination Probably the most well-known transmission channel is the theory of taste-based discrimination, proposed by Becker (1957). Physically attractive people are favored because people enjoy being or working with them more than with plain looking people. Discrimination is based upon prejudices correlated with people's personal characteristics and is rational in the sense that interactions with such a person generate a (dis)utility for the discriminator in case of positive (negative) discrimination. Belot et al. (2012) find that attractive contestants of a Dutch television game show are positively discriminated against unattractive ones in proceeding to the final stage of the show. However, this theory fails to explain why attractive people are also treated more favorably in one-shot interactions.[7]

[6]Goldman and Lewis (1977) rated the social skills during telephone calls such that the raters were not influenced by the physical appearance of the subject.

[7]In particular, attractive contestants are much more likely to reach the final round of the show, even though they are not performing better or are not more confident than unattractive contestants. Performance in this show means being the first to correctly answer trivia questions. Attractive contestants are believed to be more confident and to be more cooperative. Besides, at the end of the show a prisoner's dilemma – as we study – is played. Belot et al. (2012) also test for effects of attractiveness on cooperative behavior, however they find no significant effects. This might be caused by a too homogeneous sample of finalists from the selection of attractive contestants during

Educational attainment and self-evaluation Judge et al. (2009) provide empirical evidence that physical attractiveness positively influences educational attainment and self-evaluation and thereby the person's economic success. Related is the idea of self-fulfilling prophecies (see, e.g., Eagly et al., 1991). Physical attractiveness outshines other personality traits and links them together, such that the different treatment by others induces physically attractive people to actually behave in the way it is expected from them (see e.g., Berry and Zebrowitz, 1986).

3 Hypotheses

In line with the literature, we expect that facially attractive people do not behave differently than less facially attractive people such that they should be no more or less cooperative in the prisoner's dilemma.

Hypothesis 1. *Facially attractive contestants are no more or less likely to cooperate than less facially attractive contestants.*

But, since physical attractive people seem to be more successful in economic terms, we expect that facially attractive contestants are treated differently in the prisoner's dilemma, in particular, we expect that facially attractive contestants receive a more cooperative responding than less facially attractive people.

Hypothesis 2. *Contestants are more likely to cooperate with a facially attractive opponent than with a less facially attractive opponent.*

Studies of dating and marital choice have shown that people tend to pair off with a partner who is similar to themselves in terms of physical attractiveness (see e.g., Berscheid et al., 1971; Murstein, 1972). Therefore, one could expect that pairs of contestants who are similar in terms of facial attractiveness behave differently than contestants in mixed-pairs in the prisoner's dilemma.

Hypothesis 3. *Pairs of contestants who are similar in terms of facial attractiveness behave differently than contestants who are unsimilar in terms of facial attractiveness in the prisoner's dilemma.*

Further, we expect facially attractive contestants to obtain higher monetary gains, i.e., a monetary beauty premium. In the (modified) prisoner's dilemma, a contestant always receives a non-negative payoff, if her opponent cooperates. Presuming that Hypothesis 2 holds, facially attractive contestants should take more money home from play than less facially attractive contestants.

---

the pre-play.

Hypothesis 4. *Facially attractive contestants earn more money than less facially attractive contestants.*

One may also expect that there are differences with respect to gender and age. On the basis of studies about gender differences in cooperative behavior (e.g., Kahn et al., 1971; Ortmann and Tichy, 1999), males and females might approach the prisoner's dilemma with different concerns. However, the results are ambiguous.[8] From the literature one could, for instance, expect attractive females to be more cooperative than males (see Ortmann and Tichy, 1999), and since males tend to be more likely to do females a favor (see Ashmore and Longo, 1995), a male contestant might be more cooperative if he faces an attractive female opponent, or vice versa. Studies about differences in cooperative behavior with respect to age are very limited and often report effects on age as a byproduct of their research.[9] Thus, we have no clear-cut hypothesis concerning effects of age and gender on cooperation.

4 Data

4.1 The television game show "Golden Balls"

We collect field data from the British television game show "Golden Balls". In the final stage of the show, two contestants play for a jackpot via a slightly modified one-shot prisoner's dilemma game. Each finalist needs to simultaneously select among two choices: cooperate ("split") or defect ("steal"). Before the prisoner's dilemma game is played, the contestants have to pass two rounds of pre-play in which stakes are accumulated for the jackpot and in which the two finalists are selected.[10]
Table I presents summary statistics about the outcomes of the prisoner's dilemma game and of the contestants' personal characteristics. The unilateral cooperation

[8]In an experiment on the prisoner's dilemma game Kahn et al. (1971) find that overall male subjects tend to be more cooperative than female subjects, and that females are more cooperative when playing the game with a male opponent than with one of the same sex. Controversially, Ortmann and Tichy (1999) find that women cooperate significantly more than men in a repeated prisoner's dilemma in the laboratory, but cooperation rates equalize over time. This matches evidence from psychology that males are motivated to win and females are more concerned with interpersonal situation than with winning (see e.g., Amidjaja and Vinacke, 1965). For a survey on gender differences in preferences see Croson and Gneezy (2009).

[9]Using data from television games shows, e.g., List (2006) finds that older contestants (age $\geq$ 31) are significantly more likely to cooperate than younger contestants (age $<$ 31); whereas Belot et al. (2010) finds no age effect.

[10]The show starts with four contestants and after each pre-play round one contestant is voted to leave the game by the other contestants. In the final round the prisoner's dilemma game is played. The accumulation of the jackpot does not involve contestant's cognitive ability or an effort task. Throughout the game show contestants face each other and are allowed to communicate with each other. For a detailed description of the game show see Section 2.1 of Chapter 2.

rate is 54%, and contestants mutually cooperate (defect) in 32% (25%) of cases. 54% percent of final players are female and the mean age is 37, with an age interval from 18 to above 70. On average, the jackpot amounts to £12 912.

Table I: Summary statistics Golden Balls

Variable	Mean	Std. dev.	Min.	Max.	N
Decision variables					
Cooperate[a]	0.54	0.5	0	1	422
Mutual decision (0="steal-steal", 1="steal-split", 2="split-split")	1.07	0.76	0	2	422
Amount money taken home	4 614.8	10 799.03	0	93 250	422
Demographics					
Male	0.46	0.5	0	1	422
Age (cont.)[b]	3.21	1.1	1	6.5	422
Age of male (cont.)	3.26	1.07	1	6.4	422
Age of female (cont.)	3.19	1.09	1.25	6.6	422
White	0.94	0.25	0	1	422
London	0.11	0.32	0	1	422
England (1 = ENG, 0 = SCO, WAL, NIR, IRL)	0.85	0.36	0	1	420
Social job (reputation)[c]	0.16	0.36	0	1	421
Unexperienced (series 1)	0.19	0.39	0	1	422
Experienced (series 4)	0.19	0.39	0	1	422
Stake size					
Jackpot	12912.33	18213.95	3	93250	422
Potential jackpot[d]	50329.69	29946.46	5000	168100	422
Pre-play & communication					
Accumulated most money	0.5	0.5	0	1	422
Selected higher values in bin/win	0.5	0.5	0	1	422
Selected most killers in bin/win	0.33	0.47	0	1	422
"Should have left the game"[e]	0.26	0.44	0	1	422
Lied during pre-play	0.62	0.49	0	1	422
Promise or vow	0.42	0.494	0	1	422
Handshake	0.33	0.47	0	1	422

[a] Separated by gender and age, men (women) are found to cooperate in 51% (56%) of the observations; and older contestants ($\geq$ 37 years) cooperate more often than younger contestants ($<$ 37 years), with cooperation rates of 63% and 46%, respectively.

[b] Age is judged on a 7-item scale (see questionnaire and section 4.2), where 3="30-40", and 4="40-50" implying that the scale average of 3.21 equals a mean age of 37 years.

[c] A social job is defined as a job in which people care for other people, e.g., doctors, nurses, child minders, social workers, teachers, police officers, firemen, soldiers.

[d] Maximal amount of money the contestants potentially can gain in the prisoner's dilemma.

[e] The variable "should have left the game" points at the player who is the "weakest" in material terms in round 2.

4.2 Evaluation of facial appearance

We evaluate the contestants' facial appearance using a panel of independent raters. The raters are recruited at the Euro-Airport Basel, at the University of Zurich, and at the University of the elderly of Zurich.[11] All 844 contestants are judged by 728 raters and, from those, 365 raters judged the 422 finalists. Each rater was asked to individually rate the facial appearance of five randomly assigned contestants, of which two or three were male or female. On average, a finalist is judged by 4.3 raters.

Table II reports summary statistics for the finalists' raters and Table X in the Appendix for all contestants' raters. The mean age of the 365 raters is 41 years and 50% are male.

Table II: Summary statistics of the finalists' raters

Rater's variable	**Mean**	**Std. dev.**	**Min.**	**Max.**	**N**
Male	0.5	0.5	0	1	365
Age (in years)	40.88	15.54	17	93	361
Age of male (in years)	41.58	15.60	17	93	183
Age of female (in years)	40.15	15.50	18	86	178
Female ($\geq$ 40.15 years)	53.51	10.36	41	86	86
Female ($<$ 40.15 years)	27.66	6.38	18	40	92
Male ($\geq$ 41.58 years)	55.21	11.42	42	93	85
Male ($<$ 41.58 years)	29.76	6.28	17	41	98

The survey is questionnaire based. For an illustration of a sample questionnaire see Figure I. Each questionnaire contains two portrait photographs of the same contestant. To receive non-biased evaluations and to reduce measurement error, these photographs are all selected from the same two sequences of the game show such that the photograph showed once a neutral facial expression with a view to the camera and once a view to the side of the camera.[12]

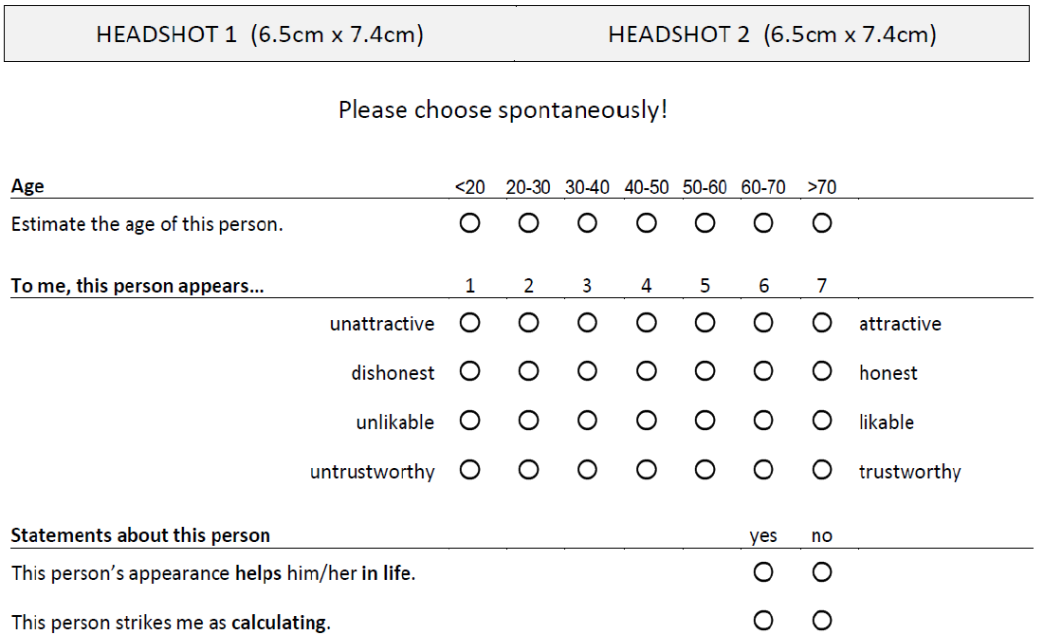
The questionnaire was divided into three parts. In the first part, raters were asked to judge the contestant with respect to her age on a 7-item scale, with the categories fitting the person either "<20", "20-30", "30-40", "40-50", "50-60", "60-70", ">70". The second part includes assessments of the contestant's appearance using 4 opposite word pairs, i.e., "attractive - unattractive", "likable - unlikable", "trustworthy - untrustworthy", and "honest - dishonest". The photographs were rated on a 1-to-7 point scale, where 1 equals very unattractive, 4 comprises a neutral position and 7

[11]The University of Zurich provides lectures for a senior audience, that are mainly attended by retired people.

[12]If possible we chose a neutral facial expression of the contestant, otherwise a positive one was chosen, but never a negative or disadvantageous one.

equals very attractive.[13] In the last part of the questionnaire we asked the raters to give a binary response (Yes/No) to the two statements "this person's appearance helps him/her in life" and "this person strikes me as calculating".

Figure I: Sample questionnaire



4.3 A measure of facial appearance

The raters' evaluations are used to construct a facial appearance measure for each of the four items: attractiveness, honesty, likability, and trustworthiness. We account for rater specific variation in the perception of the respective appearance item in the construction of each measure.[14] For each rater j we calculate the average across all images rated by her, $\bar{x}_j$. Then we mean-center the ratings by subtracting the rater's mean from each rater's individual rating of image i, x_{ij}. This demeaned ratings are

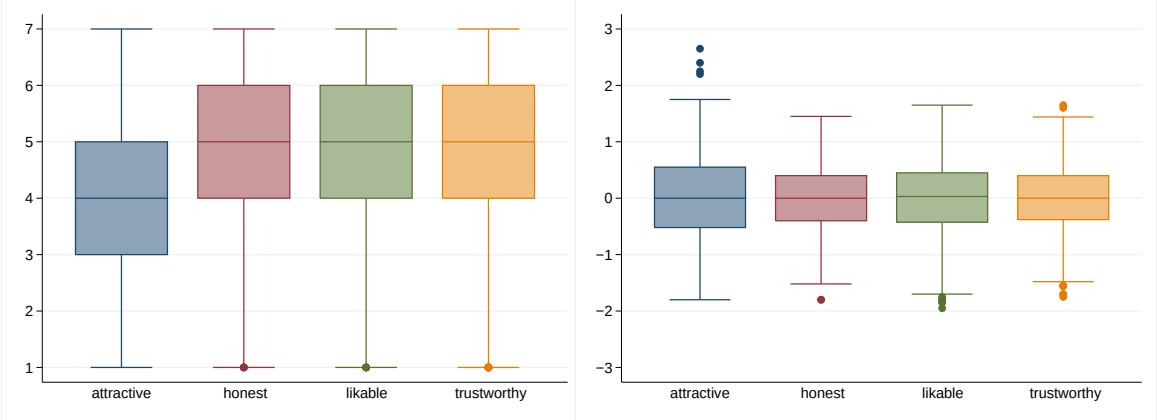
[13]The use of a 7-item scale, which includes a neutral element, allows for sufficient diversification and is standard in the relevant literature, see e.g., Alreck and Settle (1995), pp.113-114.

[14]Raters may differ in how to interpret the 7-item scale, and vary in anchoring their average rating above, below or close to 4. Student t-tests of differences of means of the attractiveness ratings by the raters' gender are $t = 2.799$, $p = 0.005$, for likability $t = 4.746$, $p = 0.000$, for trustworthiness $t = 4.913$, $p = 0.000$, and for honesty $t = 4.591$, $p = 0.000$, i.e., female raters tend to give higher ratings than male raters. This problem is also persistent with respect to the raters' age. A second problem with a direct measure of facial attractiveness, which is often addressed in the literature, is that images of females tend to be rated higher than images of males (see e.g., Jackson et al, 1995; Wilson and Eckel, 2006). Since we control directly for gender-related effects in the regressions, the attractiveness measure can be used uncorrected for gender.

now anchored at 0 for each rater, such that we have corrected for rater-specific fixed effects (see Figure II). We now take the mean of all mean-centered ratings for image i over all raters, resulting in the particular facial item measure for each image. The exact formula is as follows,

$$[\text{Facial-item}]_i = E_j[x_{ij} - \bar{x}_j] \quad \text{with} \quad \begin{cases} i = & \text{i-th contestant;} \\ j = & \text{j-th rater.} \end{cases}$$

Figure II: Boxplot [left] raw and [right] demeaned variables of facial appearance, N=422



Additionally, we construct two variables from the binary statements "appearance helps in life" and "appearance strikes as calculating" by taking the mean for each image. Table III provides summary statistics of all appearance measures.

Table III: Summary statistics facial appearance

Variable	Mean	Std. Dev.	Min.	Max.	N
Attractiveness (mean-centered, cont.)	0	0.79	-1.8	2.65	422
Honesty (mean-centered, cont.)	0	0.6	-1.8	1.45	422
Likability (mean-centered, cont.)	0	0.66	-1.95	1.65	422
Trustworthiness (mean-centered, cont.)	0	0.6	-1.75	1.65	422
Appearance helps in life (cont.)	0.58	0.3	0	1	422
Appearance strikes as calculating (cont.)	0.35	0.25	0	1	422

All four facial appearance variables are highly positively correlated (see Table IV), which is also reflected in a sufficiently high Cronbach coefficient alpha of $\alpha = 0.82$.[15]

Table IV: Correlation matrix for mean-centered variables of facial appearance

	Attractiveness	Honesty	Likability	Trustworthiness	Helps in life
Attractiveness	1				
Honesty	0.288***	1			
Likability	0.461***	0.651***	1		
Trustworthiness	0.340***	0.760***	0.680***	1	
Helps in life	0.643***	0.229***	0.389***	0.299***	1
Strikes as calculating	0.007	−0.313***	−0.288***	−0.315***	0.034

The statement variable "appearance helps in life" is also positively correlated with all four appearance variables. However, the statement variable "appearance strikes as calculating" is not correlated with attractiveness and "appearance helps in life", and is even negatively correlated with the remaining facial appearance variables.

In the following we focus on the facial attractiveness, since it is a crucial part of the first impression of a person; and it is a stable characteristic which is almost impossible to mimic. The two statement variables are used as additional controls. There are many options to define an attractive person. We use the following classifications: First, we classify a contestant as *facially attractive* if her facial attractiveness rating lies above or is equal to the mean over all facial attractiveness ratings, and as *facially unattractive* if her facial attractiveness rating lies below the mean over all facial attractiveness ratings. The average mean-centered rating of facially attractive contestants is 0.599 and the one of facially unattractive contestants is −0.650. Second, we define extreme measures of facial attractiveness. A contestant is classified as *most attractive* if her facial attractiveness rating lies within the top 10% percentile of the distribution of facial attractiveness and as *least attractive* if her facial attractiveness rating lies within the bottom 10% percentile. Contestants who are rated as most attractive receive, on average, a mean-centered rating of 1.389 and those who are rated as least attractive receive a mean-centered rating of 1.386. For an illustration of the distribution of facial attractiveness see Figure II. Third, we define a contestant's *appearance to be helpful in life (to strike as calculating)* if her "helps in life"-rating ("strikes as calculating"-rating) lies above or is equal to the mean over all "helps in life"-ratings ("strikes as calculating"-ratings), and as *not*

[15]We use Cronbach's alpha for standardized variables to measure the inter-item reliability for facial appearance. The measure adjusts for item specific mean and variance. Also a nonparametric test for testing whether samples originate from the same distribution cannot be rejected (Kruskal-Wallis $K = 0.9868$).

to be helpful in life (not to strike as calculating) if her "helps in life"-rating ("strikes as calculating"-rating) lies below this mean. Table V summarizes the binary attractiveness and statement variables. A detailed description of the distribution of attractive and unattractive finalists with respect to gender and age is provided by Table XI in the appendix.

Table V: Summary statistics facial attractiveness

Variable	Mean	Std. dev.	Min.	Max.	N
Attractiveness (mean-centered, d)	0.52	0.5	0	1	422
Most attractive (90 % percentile, d)	0.1	0.3	0	1	422
Least attractive (10 % percentile, d)	0.1	0.3	0	1	422
Helps in life (d)	0.50	0.50	0	1	422
Strikes as calculating (d)	0.47	0.50	0	1	422

5 Results

5.1 Facial attractiveness

In order to investigate the effect of facial attractiveness on cooperative behavior we use several binary probit models with the decision to cooperate as the dependent variable (with $y_i = 1$ equal cooperate; $y_i = 0$ equal defect). Throughout the analysis, we control for effects and interactions related to the contestant's gender and age, various demographic variables, as well as variables of stake size, communication, and variables describing the course of events of the game show previous to the prisoner's dilemma (pre-play), see Table I. To quantify the influence of the explanatory variables on the predicted probability to cooperate we compute marginal effects, following the method of Norten et al. (2004). See also Appendix A.3 of Chapter 2.

Own attractiveness The results depicted in Table XII in the appendix, model (1) to (4), show that facially attractive contestants do not behave differently with respect to cooperativeness than facially unattractive contestants, independent of the specification of the attractiveness measure. This finding supports Hypothesis 1. Facially attractive contestants are not more cooperative than facially unattractive contestants.[16]

[16]Table XII in the appendix reports the regression results including a dummy variable for the attractive contestant. For robustness of all our results, we also estimate the regressions including (i) the continuous measure for attractiveness, (ii) the mean over the four facial appearance variables, (iii) the predicted factors obtained in a confirmatory factor analysis, (iv) a normalized attractive-

Whereas, overall we find no difference between facially attractive and unattractive contestants, model (5) of Table XII shows that attractiveness has even a slightly negative effect on cooperative behavior when considering the extreme measure of attractiveness. If a contestant is rated as most attractive, she is actually 16.9 percentage points more likely to defect than if she is neither most attractive nor least attractive. Further, there are some qualifications with respect to gender and age.[17] Both, gender and age seem to mediate the effect of a contestant's attractiveness on cooperation, see Table XIII in the appendix. Attractive females (model (1)) and attractive younger contestants (model (3)) show a more cooperative behavior, whereas attractive males (model (1)) and attractive older contestants (model (3)) cooperate less.

Opponent's attractiveness We now turn to the impact of the opponent's facial attractiveness on a contestant's willingness to cooperate. The regression results in Table XII in the appendix, model (1) to (5), show that a contestant is $11 - 17$ percentage points more likely to cooperate when facing an attractive opponent than when facing an unattractive opponent. This result strongly supports Hypothesis 2. Attractive contestants are rewarded with greater cooperativeness, and this provides attractive contestants a beauty premium. The premium is independent of the opponent's gender and age, see Table XIII in the appendix, model (2) and (4). Furthermore, our results show that least attractive contestants suffer a beauty penalty due to lower cooperativeness towards them. As model (5) reports, a contestant is 20.4 percentage points less likely to cooperate if the opponent is rated to be least attractive than if the opponent is neither rated to be most nor least attractive. There is no significant effect on cooperative behavior if the opponent is rated to be most attractive. There is no significant effect on cooperative behavior if the opponent is rated to be most attractive. The results hold independently of the definition of the extreme measure of attractiveness. These findings suggest that contestants rather focus on the opponent's "negative" than "positive" appearance in the decision to cooperate.

---

ness measure in line with Mobius and Rosenblat (2006), in which our attractiveness measure is normalized across all contestants, (v) a normalized measure in line with Biddle and Hamermesh (1998) in which the normalization is across all raters, and (vi) the mean and median attractiveness ratings from the raw data, where the mean (median) is 0.443 (0.728), with an average rating of facially attractive contestants above the mean (median) of 5.552 (4.914), and below the mean (median) of 3.196 (2.357). All measures produce qualitatively the same results.

[17]Irrespective of a contestant's own attractiveness, we find a very strong and significant correlation between age and cooperative behavior (see Table XII in the appendix). Older contestants ($\geq$ 37 years) are much more likely to cooperate than younger contestants ($<$ 37 years), regardless of the age of the opponent. This result is in line with List (2006) who finds that contestants $\geq$ 31 years are significantly more likely to cooperate than younger contestants. Concerning gender, we find no direct effect, which is contrary to the studies of Kahn et al (1971) and Ortmann and Tichy (1999). But we find that younger males are significantly more likely to defect, and, as age increases, males are more likely to cooperate than females (see Table XIV in the appendix).

The results remain unchanged when adding the two binary statement variables, i.e., whether the "contestant's appearance helps her in life" and whether the "contestant's appearance strikes as calculating" (see model (3) and (4) of Table XII in the appendix). There is additional evidence that a contestants is less likely to cooperate if the opponent's appearance is rated as to "help her in life" than if it is not. But we find no interaction effect of the binary statement variables and our attractiveness measures, which indicates that the statement variables have not much additional explanatory power.[18]

Similarity Since contestants behave more cooperatively towards the attractive counterpart, the question arises whether pairs of attractive contestants behave differently in the prisoner's dilemma than pairs of unattractive contestants or pairs who are mixed in terms of facial attractiveness. We denote pairs of contestants as *similar team* if either both contestants are facially attractive or both are facially unattractive, and we identify pairs of contestants as an *attractive team* if both are facially attractive, and as an *unattractive team* if both are facially unattractive. Table VI reports the regression results.

Table VI: Results from binary probit regressions on unilateral cooperation, considering teams variables

	Marginal effects					
	Model (1)		Model (2)		Model (3)	
Similar team (d)	-0.043	(0.054)				
Attractive team (d)			0.045	(0.067)	0.027	(0.070)
Unattractive team (d)			-0.154**	(0.064)	-0.113*	(0.067)
Demographics	yes		yes		yes	
Stake size	yes		yes		yes	
Pre-play & communication	–		–		yes	
Wald χ^2	40.40***		46.68***		68.32***	
Log-Likelihood	-266.77		-263.49		-250.48	
Adjusted R^2	0.034		0.042		0.049	
N	419		419		419	
Number of clusters	211		211		211	

Note: Binary probit regressions of the decision either to cooperate ($y_i = 1$) or defect ($y_i = 0$) in the prisoner's dilemma game. The "team variables" are indicators and equal 1 if the team is so composed and 0 otherwise, e.g., the variable "similar team" equals 1 if both contestants are either attractive or unattractive, and 0 otherwise. The marginal effect of the respective explanatory variable determines the effective change of this variable on team i's predicted probability to cooperate ("split"). (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

We find no evidence that contestants who are similar with respect to attractiveness are more or less likely to cooperate or defect, see model (1). However, we find that

[18]Note, that the effect of the variable "the contestant's appearance helps her in life" without the inclusion of the facially attractiveness variables is positive, but not significant (table unreported).

an unattractive contestant is 11.3 to 15.4 percentage points less likely to cooperate with her unattractive counterpart compared to teams of contestants who are both attractive or mixed, see model (2) and (3). This improves our attractiveness-results: contestants not only behave more cooperatively towards an attractive partner, but also more deceitfully towards an unattractive partner, and this unattractive-penalty is likely to dominate. Addressing the mutual cooperation outcomes using ordered probits, also shows that similar teams as well as teams of attractive contestants are no more or less likely to reach a certain outcome, but that pairs of unattractive contestants are significantly less likely to manage to mutually cooperate (tables unreported).
Furthermore, we estimate regressions including the relative difference between both final contestants' attractiveness, i.e., the distance in attractiveness between both contestants, and including an index of the contestants' similarity with respect to facial attractiveness, age, and gender, weighting each component by one-third. All measures do not matter for the contestant's decision to cooperate (tables unreported). Summarizing, we cannot reject Hypothesis 3.

Thus, our results provide evidence for a (causal) relationship between the opponent's attractiveness and cooperative behavior. Facially attractive contestants are able to provoke cooperation from their counterpart, independent of their gender or age. But we do not find a significant difference in behavior between facially attractive and facially unattractive contestants.

5.2 Beauty premium

The results of the previous section should also translate into a monetary beauty premium, i.e., into higher earnings for the attractive than for the unattractive contestant. In order to quantify the marginal beauty premium, we use a standard censored tobit model (see Wooldridge, 2010, pp. 667-690). The outcome "taking no money home" from the prisoner's dilemma is interpreted as a corner solution outcome, where the response variable y_i describes the observable outcome of a contestant, which takes on the value zero with positive probability (if the opponent defects), but which is a continuous variable over strictly positive values (if the opponent cooperates). We build a (index) model around the latent variable y_i^*,

$$y_i^* = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i = \mathbf{X_i}\beta + \varepsilon_i, \quad \varepsilon_i \sim \text{NID}(0, \sigma^2),$$

where $x_1, ..., x_k$ are demographic player characteristics as well as the log value of the stake size, β denotes the response coefficient vector, including a constant, and ε_i is the random error that is normally, independently, and identically distributed (NID), with σ^2 denoting the variance. The latent variable y_i^* is observed whenever

it is positive, and is censored at zero otherwise,

$$y_i = \begin{cases} y_i^*, & \text{if } y_i^* > 0; \\ 0, & \text{otherwise.} \end{cases}$$

The conditional expectation, $E[y_i|\mathbf{X_i}]$, that contestant i receives y_i is determined as

$$E[y_i|\mathbf{X_i}] = Pr(y_i > 0|\mathbf{X_i})E[y_i|\mathbf{X_i}, y_i > 0] = \Phi\left(\frac{\mathbf{X_i}\beta}{\sigma}\right)\mathbf{X_i}\beta + \sigma\phi\left(\frac{\mathbf{X_i}\beta}{\sigma}\right),$$

with $\Phi(\cdot)$ the standard normal cumulative distribution function, and $\phi(\cdot)$ the standard normal density function. The marginal effect for the j-th independent (continuous) variable is computed as

$$ME_j = \frac{\partial E[y_i|\mathbf{X_i}]}{\partial x_j} = \Phi\left(\frac{\mathbf{X_i}\beta}{\sigma}\right)\beta_j, \quad j = 2, ..., K,$$

where the estimated scale factor $\Phi\left(\frac{\mathbf{X_i}\hat{\beta}}{\hat{\sigma}}\right)$ is the estimated probability, $\hat{Pr}(y_i > 0|\mathbf{X_i})$, of observing a positive response given $\mathbf{X_i}$. Table VII reports the regression results.

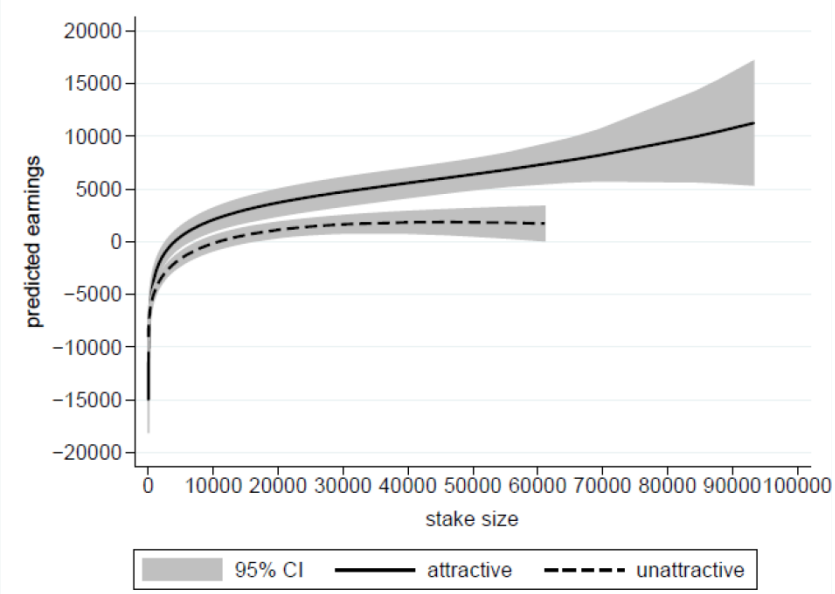
Table VII: Censored tobit regression results for $E[y_i|\mathbf{X_i}]$

	Marginal effects			
	Model (1)		Model (2)	
Attractiveness (standardized)	1783.681*	(1.83)	2152.525**	(2.26)
Opp. attractiveness (standardized)			1327.387	(1.60)
Log(jackpot)	1993.899***	(4.54)	1788.198***	(4.42)
Demographics	yes		yes	
Opp. demographics	–		yes	
F-Statistic	3.19***		2.84***	
Log-Likelihood	-2593.35		-2578.46	
Adjusted R^2	0.003		0.006	
$\hat{\sigma}$	15059.81		14483.33	
N	419		419	
Number of clusters	211		211	

Note: Censored tobit regression for the conditional expectation, $E[y_i|\mathbf{X_i}]$, that contestant i receives the outcome y_i. y_i takes on the value zero with positive probability if the opponent defects, and is a continuous variable over strictly positive values if the opponent cooperates. Marginal effects of $E[y_i|\mathbf{X_i}]$ are reported to quantify the expected increase in earnings. (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Model (1) includes the contestant's facial attractiveness and demographic characteristics. Model (2) includes in addition the opponent's facial attractiveness and demographic characteristics. We also include the logarithm of the stake size in order to control for different stake levels. We find that a one standard deviation increase in facial attractiveness increases a contestant's expected earnings by £1 784 (£2 153) depending on the controls included.[19] Figure III provides an illustration of the beauty premium based on model (2) of Table VII, and depicts the predicted earnings for attractive and unattractive contestants as a function of stake size. The figure shows that the predicted earnings of facially attractive contestants are always above the earnings of unattractive contestants.

Figure III: Predicted earnings by attractive and unattractive contestants



Further, we estimate the effect of the contestant's expected earnings conditional on being positive, $E[y_i|\mathbf{X_i}, y_i > 0]$, as well as the contestant's probability of obtaining positive earnings, $Pr(y_i > 0|\mathbf{X_i})$, separately.
A contestant i's expected earnings conditional on being positive are given by

$$E[y_i|\mathbf{X_i}, y_i > 0] = \mathbf{X_i}\beta + \sigma \left[\frac{\phi(\mathbf{X_i}\beta/\sigma)}{\Phi(\mathbf{X_i}\beta/\sigma)} \right],$$

where the term $\frac{\phi(\mathbf{X_i}\beta/\sigma)}{\Phi(\mathbf{X_i}\beta/\sigma)}$ denotes the inverse Mills ratio evaluated at $\frac{\mathbf{X_i}\beta}{\sigma}$. The marginal effect for the j-th independent (continuous) variable on $E[y_i|\mathbf{X_i}]$ is com-

[19]The standardization of the facial attractiveness variable allows us to interpret the regression coefficient as the effect of a one-standard deviation increase (decrease) in facial attractiveness.

puted as

$$ME_j = \frac{\partial E[y_i|\mathbf{X_i}]}{\partial x_j} = \beta_j \left[1 - \frac{\phi(\mathbf{X_i}\beta/\sigma)}{\Phi(\mathbf{X_i}\beta/\sigma)} \left(\frac{\mathbf{X_i}\beta}{\sigma} + \frac{\phi(\mathbf{X_i}\beta/\sigma)}{\Phi(\mathbf{X_i}\beta/\sigma)}\right)\right].$$

Concerning the marginal effect for the j-th independent (continuous) variable on the probability of observing positive earnings given $\mathbf{X_i}$, $Pr(y_i > 0|\mathbf{X_i}) = \Phi(\mathbf{X_i}\beta/\sigma)$, we have

$$\frac{\partial Pr(y_i > 0|\mathbf{X_i})}{\partial x_j} = \frac{\beta_j}{\sigma}\phi\left(\frac{\mathbf{X_i}\beta}{\sigma}\right).$$

Table XV and Table XVI in the appendix report the regression results of the marginal effects, respectively. The marginal effect for $E[y_i|\mathbf{X_i}, y_i > 0]$ amounts to £617 (model (1)) and £741 (model (2)), depending on the number of controls included, and is smaller in size compared to the marginal effect for $E[y_i|\mathbf{X_i}]$. The probability of obtaining positive earnings increases by 4.7 (5.9) percentage points. To sum up, we find strong evidence in favor of Hypothesis 4, i.e., facially attractive contestants earn a beauty premium in the prisoner's dilemma game.

6 Transmission channels

Our results show that attractive people are able to provoke cooperative behavior from their opponent and, since they are not more or less cooperative than unattractive people, they obtain a beauty premium in the prisoner's dilemma. In this section we will address the potential transmission channels suggested by the literature (see Section 2.2) and evaluate their explanatory power for our observed effects of attractiveness.

Beauty-is-good stereotyping and taste-based discrimination According to the most prominent theory called "beauty-is-good" stereotyping, effects of attractiveness on behavior are caused by stereotype believes about attractive people. People are said to believe that attractive ones behave more pro-socially than less attractive ones. In the presence of interdependent social preferences, people holding this belief may then behave more cooperatively towards attractive people with the intention to reciprocate cooperation. Hence, these stereotype beliefs can mediate people's behavior. Recall, that we observe a higher likelihood of cooperation towards attractive contestants in the prisoner's dilemma. If the theory of "beauty-is-good" stereotyping is driving our results, this effect should be independent of the selected sample. However, theories from evolutionary psychology argue that effects of physical attractiveness on behavior originate in primeval partner selection and

therefore predict the effects to be more prevalent in mixed-gender interactions (see e.g., Cosmides and Tooby, 1987). To scrutinize this argument, we parse the data in mixed-gender and same-gender interactions (63% and 37% of all interactions). The two subsamples are not different regarding the observed cooperation rates (54% in mixed- vs. 53% in same-gender interactions). Within the two subsamples, we run several probit regressions to evaluate the influence of facial attractiveness on a contestant's propensity to cooperate, controlling for demographic characteristics and stake size.

Table VIII: Results from binary probit regressions of the decision in the prisoner's dilemma in mixed- and same-gender interactions

	Marginal effects			
	(1) Mixed-gender		(2) Same-gender	
Attractiveness (mean-centered, d)	0.054	(0.068)	0.127	(0.089)
Opp. attractiveness (mean-centered, d)	0.142**	(0.063)	0.040	(0.089)
Demographics	yes		yes	
Stake size	yes		yes	
Wald χ^2	45.15***		19.66*	
Log-Likelihood	-161.34		-97.00	
Adjusted R^2	0.047		-0.032	
N	265		154	
Number of clusters	133		78	

Note: Binary probit regressions of the decision either to cooperate ($y_i = 1$) or to defect ($y_i = 0$) in the prisoner's dilemma game, restricting the sample to (1) mixed-gender and (2) same-gender interactions. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to cooperate ("split"). Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

As the regression results of Table VIII show, attractiveness only matters in mixed-gender interactions, but not in same-gender ones. Males and females are about 14 percentage points more likely to cooperate if they face a facially attractive opponent of the other sex than if they face an unattractive opponent.[20] The finding applies across sexes, both females and males are biased towards attractiveness if facing someone of the other sex.[21]

Further, we find no effect of attractiveness in the two group-decisions taken by

[20]This result is also displayed in the fact that facially attractive contestants enjoy a beauty premium in mixed-gender interactions, but not in same-gender ones (tobit regression tables unreported). A one standard deviation increase in attractiveness results in significantly higher gains for the attractive in mixed-gender ($p = 0.003$), but not in same-gender interactions ($p = 0.840$).

[21]This extends the finding of Ashmore and Longo (1995), who note that only attractive females tend to have the ability to make males more likely to do them a favor. We show that males have the same ability.

the contestants prior to the prisoner's dilemma. Contestants might favor to be in the final round with a facially attractive person. If facially attractive contestants are more likely to be in the final than unattractive contestants, then taste-based discrimination (Becker, 1957) could explain this beauty premium. Using binary and ordered probit models, we estimate whether attractive contestants are more or less likely to be voted off the game and whether attractiveness affects the likelihood to receive a certain number of votes after the first and second round of pre-play. We find no significant effect of attractiveness on the voting outcome (tables unreported). This is also reflected in an almost equal share of attractive contestants in the final and initial round of the game show (52.4% vs. 51.8%), see Table IX below.[22]

Table IX: Distribution of players above average attractiveness per round

	Number of players	**Avg. number of attractive players**	**Avg. share of attractive players**
Round 1	4	2.07	51.78%
Round 2	3	1.59	52.92%
Final (PD)	2	1.05	52.37%

The absence of a beauty premium or plainness penalty in the pre-play as well as in same-gender interactions in the final suggests that neither beauty-is-good stereotyping nor taste-based discrimination can explain our results consistently. It rather seems that the ability of attractiveness to mediate behavior is entangled to the opponent's sex and that people have a preference to cooperate with someone to whom they are personally attracted.

Vocal attractiveness and social skills In addition to the theories discussed above, the literature reports that physical and vocal attractiveness are highly correlated and thus suggests the existence of a vocal rather than a visual attractiveness stereotype (see, e.g., Zuckerman and Driver, 1989). Furthermore, physically attractive people are also assigned stronger verbal and social skills (e.g., Goldman and Lewis, 1977). In the presence of other regarding preferences, vocal attractiveness and potentially strong verbal and social skills might enable attractive people to trigger cooperative behavior off their opponents. Although our data does not allow us to directly test for the impact of (perceived) vocal attractiveness on cooperative behavior, we can indirectly test for a correlation between facial attractiveness

[22]In all regressions we control for demographic characteristics such as gender, age, race and place of residence as well as for objective voting criteria such as the stake size a contestant accumulated, and whether a contestant lied in a previous round of the pre-play. We also follow Belot et al. (2008) approach and rank the contestants by their facial attractiveness to explain the likelihood to be voted off during the game, which yields no significant effects either (table unreported). Our finding is in contrast to Belot et al. (2008), who find that attractive people are positively discriminated against unattractive people.

and communication which comprises verbal and social skills. Shortly before the prisoner's dilemma is played both contestants are given some extra time to talk to each other. In these short conversations (on average, 38 seconds) each contestant tries to convince her opponent to cooperate. Since it has been shown that promises effect people's behavior in experiments (e.g., Charness and Dufwenberg, 2006, and Vanberg, 2008) and in the field (e.g., Belot et al., 2010), we code whether a contestant explicitly promises her opponent to cooperate.[23] Furthermore, we observe that contestants use handshakes to corroborate their mutual intention to cooperate and therefore we also code whether two contestants shake hands. We find that promises have a significantly positive impact on cooperative behavior, whereas handshakes have a negative one. These effects are robust to various specifications of the regression model. The effect of attractiveness on cooperation remains almost unchanged when we add the variables of communication as controls. We find also no evidence that the effects of attractiveness and communication are interacted. However, we find limited evidence that contestants are significantly more likely to state a promise if the opponent is attractive (table unreported).[24] Even though we cannot entirely exclude that promises underly the effects of attractiveness, the theory of better social and verbal skills or vocal attractiveness can also not explain the absence of the beauty premium in the pre-play.

7 Conclusion

This study shows that men and women are biased by attractiveness in their economic decision making when facing someone attractive of the opposite sex. We analyze the relationship between attractiveness and cooperative behavior in a high-stakes field setting with two-sided communication. Two independent data sets are combined. One on cooperation, collected from decisions (either to "split" or "steal" a jackpot) made in a slightly modified prisoner's dilemma played in the final round of a television game show. The other one on the physical attractiveness of the game

[23]We count all statements as a promise when they contain either the word "promise" or "swear" or they are a statement of intent. Examples are "I promise to split", "I promise I will not steal", "I swear I will split", "I swear I will not steal", "I will split", or "I will not steal".

[24]If attractive contestants are better in terms of verbal and social skills, a promise or handshake of a facially attractive contestant might be more convincing than a promise or handshake of a facially unattractive contestant. Further, we could expect that facially attractive contestants are more likely to elicit a promise from their opponent and less likely to engage in a handshake and thereby provoke more cooperative behavior from their counterpart. Testing for interaction effects between facial attractiveness and the communication variables, reveals no additional effects (table unreported). Using binary probit regressions on the contestant's propensity to promise or to shake hands, we find that facially attractive contestants are not more likely to state a promise or to shake hands than facially unattractive contestants (table unreported).

show's contestants using a sample of independent third-party raters.
Our results show a strong and robust effect of attractiveness: contestants are significantly more likely to cooperate with a facially attractive opponent. But facially attractive contestants are not more cooperative than facially unattractive contestants. Hence, attractive contestants are rewarded by a beauty premium, which, at the margin, amounts to up to £2 153 for an increase in attractiveness by a one standard deviation. The attractiveness effect is robust to demographic characteristics, including gender, stake size, and communication. However, the effects of attractiveness might be amplified by the attractive contestant's ability to talk their opponent into promising to cooperate, which has a significantly positive effect on a contestant's likelihood to cooperate.
Decomposing the data into same- and mixed-gender interactions reveals a new insight on effects of attractiveness. The ability of attractive contestants to elicit cooperative behavior from their opponent vanishes in interactions between two contestants of the same sex. That is, contestants are only biased by the facial attractiveness of their opponent when the opponent is of the other sex. This suggests that stereotype beliefs about attractive people, such as them being more pro-social, cannot explain our results. Since for this explanation to hold, the effect should prevail in all interactions. In line with our finding are rather theories from evolutionary psychology arguing that effects of physical attractiveness originate in primeval partner selection and should therefore be only or at least more present in mixed-gender interactions. In addition, the absence of a beauty premium or plainness penalty in the pre-play suggests that physical appearance is particularly important as soon as people are lacking objective information. Hence, we propose a preference-based mechanism as the underlying transmission channel of attractiveness on cooperative behavior. People are more likely to cooperate with someone towards whom they are personally attracted.
Our results are relevant and applicable to one-shot face-to-face interactions and are particularly important when objective information is scarce. Such situations are, for instance, job interviews or negotiations. Further, the finding that attractive people fare better in the labor market might be reinforced by the fact that attractive people benefit from greater cooperativeness towards them.

Acknowledgements

The authors thank Armin Schmutzler, Michelle Sovinsky, Ulrich Kaiser, and the seminar participants at the University of Zurich for helpful discussions and suggestions, as well as Pascal Kappeler, Holger Scriba, and Christina Richard for excellent research assistance. Financial support of the Swiss National Science Foundation is gratefully acknowledged. The data were provided to the authors by the television show producers, courtesy of Endemol UK plc, in May 2009.

Appendix

Table X: Summary statistics all raters

Variable	**Mean**	**Std. dev.**	**Min.**	**Max.**	**N**
Age (in years)	41.76	18.46	17	93	720
Age of male (in years)	39.57	17.70	17	93	371
Age of female (in years)	44.10	19.00	18	86	349
Male	0.51	0.5	0	1	728
Female ($\geq$ 44.1 years)	62.07	10.10	45	86	162
Female ($<$ 44.1 years)	28.52	7.84	18	44	187
Male ($\geq$ 39.6 years)	58.02	11.88	40	93	153
Male ($<$ 39.6 years)	26.62	5.17	17	39	218

Table XI: Distribution of attractive and unattractive finalists

	Facially attractive	Facially unattractive
Male ($<$ 37 years)	21%	27%
Male ($\geq$ 37 years)	15%	30%
Female ($<$ 37 years)	41%	18%
Female ($\geq$ 37 years)	23%	25%
	100% (N=221)	100% (N=201)

Table XII: Results from binary probit regressions on unilateral cooperation

	Marginal effects									
	Model (1)		Model (2)		Model (3)		Model (4)		Model (5)	
Attractiveness (mean-centered, d)	0.067	(0.049)	0.067	(0.052)	0.049	(0.058)	0.026	(0.060)		
Opp. attractiveness (mean-centered, d)	0.112**	(0.046)	0.125**	(0.049)	0.170***	(0.056)	0.162***	(0.057)		
Appearance helps in life (d)					0.044	(0.062)	0.062	(0.066)	0.086	(0.061)
Opp. appearance helps in life (d)					-0.091	(0.061)	-0.110*	(0.064)	-0.037	(0.059)
Appearance strikes as calculating (d)					0.010	(0.054)	0.017	(0.056)	0.011	(0.056)
Opp. appearance strikes as calculating (d)					0.016	(0.055)	-0.001	(0.056)	0.001	(0.056)
Most attractive (90% percentile, d)									-0.169*	(0.090)
Opp. most attractive (90% percentile, d)									-0.140	(0.086)
Least attractive (10% percentile, d)									-0.049	(0.091)
Opp. least attractive (10% percentile, d)									-0.204**	(0.083)
Male (d)	-0.036	(0.048)	-0.046	(0.050)	-0.039	(0.051)	-0.052	(0.056)	-0.065	(0.051)
Age (cont.)	0.081***	(0.023)	0.082***	(0.025)	0.084***	(0.026)	0.093***	(0.030)	0.084***	(0.028)
Opp. male (d)							-0.008	(0.056)		
Opp. age (cont.)							0.016	(0.028)		
Demographics	–		yes		yes		yes		yes	
Stake size	–		yes		yes		yes		yes	
Pre-play & communication	–		–		–		yes		yes	
Wald χ^2	18.59***		43.01***		47.27***		69.21***		66.6456***	
Log-Likelihood	-282.75		-264.39		-262.94		-248.53		-246.47	
Adjusted R^2	0.013		0.039		0.030		0.042		0.049	
N	422		419		419		419		419	
Number of clusters	211		211		211		211		211	

Note: Binary probit regressions of the decision either to cooperate ($y_i = 1$) or defect ($y_i = 0$) in the prisoner's dilemma game. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to cooperate ("split"). (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table XIII: Results from binary probit regressions on unilateral cooperation, including interaction terms "attractiveness*gender" and "attractiveness*age"

	Marginal effects							
	Model (1)		Model (2)		Model (3)		Model (4)	
Attractiveness*male (d)	-0.153*	(0.104)						
Opp. attractiveness*male (d)			0.099	(0.103)				
Attractiveness*age (cont.)					-0.079**	(0.044)		
Opp. attractiveness*age (cont.)							-0.011	(0.044)
Attractiveness (mean-centered, d)	0.133*	(0.078)			0.334**	(0.150)		
Opp. attractiveness (mean-centered, d)			0.068	(0.073)			0.154	(0.161)
Male (d)	0.044	(0.072)			-0.036	(0.050)	-0.054	(0.050)
Opp. male (d)			-0.055	(0.075)				
Age (cont.)	0.082***	(0.027)	0.075***	(0.024)	0.123***	(0.033)		
Opp. age (cont.)							0.006	(0.033)
Demographics	yes		yes		yes		yes	
Stake size	yes		yes		yes		yes	
Wald χ^2	36.01***		44.63***		43.93***		35.31***	
Log-Likelihood	-266.08		-265.29		-265.65		-270.02	
Adjusted R^2	0.033		0.035		0.034		0.019	
N	419		419		419		419	
Number of clusters	211		211		211		211	

Note: Binary probit regressions of the decision either to cooperate ($y_i = 1$) or defect ($y_i = 0$) in the prisoner's dilemma game. Model (1) includes the interaction between the contestant's own attractiveness and gender, whereas model (2) includes the interaction between the opponent's attractiveness and the opponent's gender. Analogously, model (3) and (4) include the interaction between the contestant's own attractiveness and age, and the opponent's attractiveness and the opponent's age, respectively. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to cooperate ("split"). (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table XIV: Results from binary probit regressions on unilateral cooperation, including the interaction term "gender*age"

	Marginal effects									
	Model (1)		Model (2)		Model (3)		Model (4)		Model (5)	
Attractiveness (mean-centered, d)			0.062	(0.052)	0.048	(0.058)	0.025	(0.059)		
Opp. attractiveness (mean-centered, d)			0.124**	(0.049)	0.168***	(0.056)	0.161***	(0.057)		
Appearance helps in life (d)					0.036	(0.062)	0.056	(0.065)	0.078	(0.061)
Opp. appearance helps in life (d)					-0.090	(0.061)	-0.108*	(0.064)	-0.036	(0.060)
Appearance strikes as calculating (d)					0.011	(0.054)	0.018	(0.057)	0.012	(0.056)
Opp. appearance strikes as calculating (d)					0.018	(0.055)	0.001	(0.056)	0.003	(0.057)
Most attractive (90% percentile, d)									-0.181**	(0.091)
Opp. most attractive (90% percentile, d)									-0.132	(0.088)
Least attractive (10% percentile, d)									-0.058	(0.091)
Opp. least attractive (10% percentile, d)									-0.208**	(0.084)
Male*age (cont.)	0.071**	(0.042)	0.067*	(0.042)	0.065*	(0.042)	0.047	(0.043)	0.057*	(0.043)
Male (d)	-0.304**	(0.147)	-0.286*	(0.149)	-0.273*	(0.149)	-0.232	(0.165)	-0.285*	(0.163)
Age (cont.)	0.037	(0.033)	0.048	(0.034)	0.050	(0.034)	0.066*	(0.039)	0.051	(0.038)
Opp. male (d)							-0.004	(0.056)		
Opp. age (cont.)							0.015	(0.028)		
Demographics	yes		yes		yes		yes		yes	
Stake size	yes		yes		yes		yes		yes	
Pre-play & communication	–		–		–		yes		yes	
Wald χ^2	40.35***		47.21***		48.56***		69.17***		68.0140***	
Log-likelihood	-266.43		-263.10		-261.74		-247.87		-245.50	
Adjusted R^2	0.035		0.040		0.030		0.040		0.049	
N	419		419		419		419		419	
Number of clusters	211		211		211		211		211	

Note: Binary probit regressions of the decision either to cooperate ("split": $y_i = 1$) or defect ("steal": $y_i = 0$) in the prisoner's dilemma game. Here, the interaction between gender and age is included as a control variable. The marginal effect of the respective explanatory variable determines the effective change of this variable on player i's predicted probability to "split". (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$).

Table XV: Censored tobit regressions results for $E[y_i|\mathbf{X_i}, y_i > 0]$

	Marginal effects			
	Model (1)		Model (2)	
Attractiveness (standardized)	616.927*	(334.835)	741.419**	(325.057)
Opp. attractiveness (standardized)			457.207	(285.004)
Log(jackpot)	689.636***	(157.020)	615.929***	(146.146)
Demographics	yes		yes	
Opp. demographics	–		yes	
F-Statistic	3.19***		2.84***	
Log-Likelihood	-2593.35		-2578.46	
Adjusted R^2	0.003		0.006	
$\hat{\sigma}$	15059.81		14483.33	
N	419		419	
Number of clusters	211		211	

Note: Censored tobit regression for the conditional expectation, $E[y_i|\mathbf{X_i}, y_i > 0]$, that player i receives a positive earnings $y_i > 0$ from the prisoner's dilemma game. Marginal effects of $E[y_i|\mathbf{X_i}, y_i > 0]$ are reported to quantify the expected increase in earnings, conditional on the earnings being positive. (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table XVI: Censored tobit regressions results for $Pr(y_i > 0|\mathbf{X}_i)$

	Marginal effects			
	Model (1)		Model (2)	
Attractiveness (standardized)	0.047*	(0.024)	0.059**	(0.025)
Opp. attractiveness (standardized)			0.036	(0.023)
Log(Jackpot)	0.053***	(0.011)	0.049***	(0.011)
Demographics	yes		yes	
Opp. demographics	–		yes	
F-Statistic	3.19***		2.84***	
Log-Likelihood	-2593.35		-2578.46	
Adjusted R^2	0.003		0.006	
$\hat{\sigma}$	15059.81		14483.33	
N	419		419	
Number of clusters	211		211	

Note: Censored tobit regression for the probability, $Pr(y_i > 0|\mathbf{X}_i)$, of obtaining positive earnings from the prisoner's dilemma game. Marginal effects of $Pr(y_i > 0|\mathbf{X}_i)$ are reported to quantify the probability increase of obtaining positive earnings. (d) for discrete change of dummy variable from 0 to 1. Standard errors are reported in parentheses and are corrected for episode clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Alreck, P. L., and R. B. Settle (1995): *The Survey Research Handbook*. Chicago: Irwin.

Amidjaja, I. R., and W. E. Vinacke (1965): "Achievement, Nurturance, and Competition in Male and Female Triads," *Journal of Personality and Social Psychology*, 2, 447–451.

Anderson, C., O. P. John, D. Keltner, and A. M. Kring (2001): "Who Attains Social Status? Effects of Personality and Physical Attractiveness in Social Groups," *Journal of Personality and Social Psychology*, 81, 116–132.

Andreoni, J., and R. Petrie (2008): "Beauty, Gender and Stereotypes: Evidence from Laboratory Experiments," *Journal of Economic Psychology*, 29, 73–93.

Antonakis, J., and O. Dalgas (2009): "Predicting Elections: Child's Play!," *Science*, 323, 1183.

Ashmore, R. D., and L. C. Longo (1995): "Accuracy of Stereotypes: What Research on Physical Attractiveness can Teach us," in *Stereotype Accuracy: Toward Appreciating Group Differences*, ed. by Y. Lee, L. Jussim, and C. McCauley, chap. 3, pp. 63–86. American Psycological Association.

Becker, G. (1957): *The Economics of Discrimination*. University of Chicago Press.

Belot, M., V. Bhaskar, and J. van de Ven (2008): "Beauty and the Sources of Discrimination," *Journal of Human Resources, forthcoming*.

——— (2010): "Promises and Cooperation: Evidence from a TV Game Show," *Journal of Economic Behavior and Organizations*, 73, 396–405.

Berggren, N., H. Jordahl, and P. Poutvaara (2010): "The Looks Of A Winner: Beauty And Electoral Success," *Journal of Public Economics*, 94, 8–15.

Berry, D. S., and L. A. Z. McArthur (1986): "Perceiving Character in Faces: The Impact of Age-Related Craniofacial Changes on Social Perception," *Psychological Bulletin*, 100(1), 3–18.

Berry, D. S., and L. Z. McArthur (1985): "Some components and consequences of a babyface," *Journal of Personality and Social Psychology*, 48, 312–323.

Berscheid, E., K. Dion, E. Walster, and G. W. Walster (1971): "Physical Attractiveness and Dating Choice: A Test of the Matching Hypothesis," *Journal of Experimental Social Psychology*, 7(2), 173–189.

Biddle, J. E., and D. S. Hamermesh (1998): "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre," *Journal of Labor Economics*, 16(1), 172–201.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74(6), 1579–1601.

Cosmides, L., and J. Tooby (1987): "From Evolution to Behavior: Evolutionary Psychology as the Missing Link," in *The Lates on the Best: Essays on Evolution and Optimality*, ed. by J. Dupré, chap. 13, pp. 277–306. Cambridge: MIT Press.

Croson, R., and U. Gneezy (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47(2), 448–447.

Dion, K., E. Berscheid, and E. Walster (1972): "What Is Beautiful Is Good," *Journal of Personality and Social Psychology*, 24, 285–290.

Eagly, A. H., R. D. Ashmore, M. G. Makhijani, and L. C. Longo (1991): "What is Beautiful is Good, But...: A Meta-Analytic Review of Research on the Physical Attractiveness Stereotype," *Psychological Bulletin*, 110(1), 109–128.

Eckel, C. C., and R. Petrie (2011): "Face Value," *American Economic Review*, 101, 1497–1513.

Erwin, P. G., and A. Calev (1984): "Beauty: More than Skin Deep?," *Journal of Social and Personal Relationships*, 1, 359–361.

Feingold, A. (1992): "Good-Looking People Are Not What We Think," *Psychological Bulletin*, 111(2), 304–341.

Goldman, W., and P. Lewis (1977): "Beautiful is Good: Evidence that the Physically Attractive are More Socially Skillful," *Journal of Experimental Social Psychology*, 13, 125–130.

Hamermesh, D. S. (2011): *Beauty Pays – Why Attractive People Are More Successful*. Princeton, New Jersey: Princeton University Press.

Hope, D. A., and J. A. Mindell (1994): "Global Social Skill Ratings: Measures of Social Behavior or Physical Attractiveness?," *Behavioral Research and Therapy*, 32, 463–469.

Jackson, L. A., J. H. Hunter, and C. N. Hodge (1995): "Physical Attractiveness and Intellectual Competence: A Meta-Analysis Review," *Social Psychology Quarterly*, 58(2), 108–122.

Judge, T. A., C. Hurst, and L. S. Simon (2009): "Does it Pay to Be Smart, Attractive, or Confident (or All Three)? Relationships Among General Mental Ability, Physical Attractivness, Core Self-Evaluation, and Income," *Journal of Applied Psychology*, 94, 742–755.

Kahn, A., J. Hottes, and W. L. Davis (1971): "Cooperation and Optimal Responding in the Prisoner's Dilemma Game: Effects of Sex and Physical Attractiveness," *Journal of Personality and Social Psychology*, 17(3), 267–279.

Kanazawa, S. (2011): "Intelligence and Physical Attractiveness," *Intelligence*, 39, 7–14.

Langlois, J. H., L. Klakanis, A. J. Rubenstein, A. Larson, M. Hallam, and M. Smoot (2000): "Maxims or Myths of Beauty? A Meta-Analysis and Theoretical Review," *Pssychological Bulletin*, 126(3), 390–423.

List, J. A. (2006): "Friend or Foe? A Natural Experiment of the Prisoner's Dilemma," *The Review of Economics and Statistics*, 88(3), 463–471.

Mobius, M. M., and T. S. Rosenblat (2006): "Why Beauty Matters," *The American Economic Review*, 96(1), 222–235.

Mueller, U., and A. Mazur (1997): "Facial Diominance in Homo Sapiens as Hinest Signaling of Male Quality," *Behavioral Ecology*, 8, 569–579.

Mulford, M., J. Orbell, C. Shatto, and J. Stockard (1998): "Physical Attractiveness, Opportunity and Success in Everyday Exchange," *The American Journal of Sociology*, 103(6), 1565–1593.

Murstein, B. I. (1972): "Physical Attractiveness and Marital Choice," *Journal of Personality and Social Psychology*, 22(1), 8–12.

Norton, E. C., H. Wang, and C. Ai (2004): "Computing Interaction Effects and Standard Errors in Logit and Probit Models," *The Stata Journal*, 4(2), 154–167.

O'Neill, J. (2003): "The Gender Gap in Wages, Circa 2000," *American Economic Review*, 93(2), 309–314.

Ortmann, A., and L. K. Tichy (1999): "Gender differences in the laboratory: evidence from prisoner's dilemma games," *Journal of Economic Behavior and Organization*, 39, 327–339.

Ottati, V. C., and M. Deiger (2002): "Visual Cues and the Candidate Evaluation Process," in *Social Psychology of Politics*, ed. by V. C. Ottati, and R. S. Tindale, pp. 75–87. New York: Kluwer Academic/Plenum.

Rosar, U., M. Klein, and T. Beckers (2008): "The Frog Pond Beauty Contest: Physical Attractiveness and Electoral Success of the Constituency Candidates at the North Rhine-Westphalia State Election of 2005," *European Journal of Political Research*, 47, 64–79.

Rosenblat, T. S. (2008): "The Beauty Premium: Physical Attractiveness and Gender in Dictator Games," *Negotiation Journal*, 24(4), 465–481.

Ruffle, B. J., and Z. Shtudiner (2010): "Are Good-Looking People More Employable?," *Working Paper*.

Solnick, S. J., and M. E. Schweitzer (1999): "The Influence of Physical Attractiveness and Gender on Ultimatum Game Decisions," *Organizational Behavior and Human Decision Process*, 79(3), 199–215.

Stirrat, M., and D. I. Perrett (2010): "Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness," *Psychological Science*, 21(3), 349–354.

Vanberg, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," *Econometrica*, 76(6), 1467–1480.

Wilson, R. K., and C. C. Eckel (2006): "Judging a Book by its Cover: Beauty and Expectations in the Trust Game," *Political Research Quarterly*, 59(189), 188–202.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd edn.

Zebrowitz, L. A., J. A. Hall, N. A. Murphy, and G. Rhodes (2002): "Looking Smart and Looking Good: Facial Cues to Intelligence and their Origins," *Personality and Social Psychology Bulletin*, 28, 238–249.

Zuckerman, M., and R. E. Driver (1989): "What Sounds Beautiful is Good: The Vocal Attractiveness Stereotype," *Journal of Nonverbal Behavior*, 13, 67–82.

Zuckerman, M., H. Hodgins, and K. Miyake (1990): "The Vocal Attractiveness Stereotype: Replication and Elaboration," *Journal of Nonverbal Behavior*, 14, 97–112.

Curriculum Vitae

Born September 2nd, 1982 in Frechen, Germany

1993 – 2002	Secondary school in Dormagen, Germany Degree: Abitur
2002 – 2007	Studies in economics and mathematics at the University of Bonn, Germany Degree: Dipl. Volkswirtin Thesis supervisor: *Prof. Dr. h.c. mult. Martin Hellwig*
2003 – 2007	Student research assistant at the Max-Planck Institute for Research on Collective Goods in Bonn, Germany
2006 – 2007	Studies in economics at the New University of Lisbon, Portugal
2008 – 2012	Doctoral studies in economics at the University of Zurich, Switzerland
2008 – 2012	Research and teaching assistant at the chair of *Prof. Dr. Armin Schmutzler* at the Department of Economics of the University of Zurich, Switzerland
2009 – 2010	Swiss program for beginning doctoral students in economics at the Study Center Gerzensee - Foundation of the Swiss National Bank, Switzerland